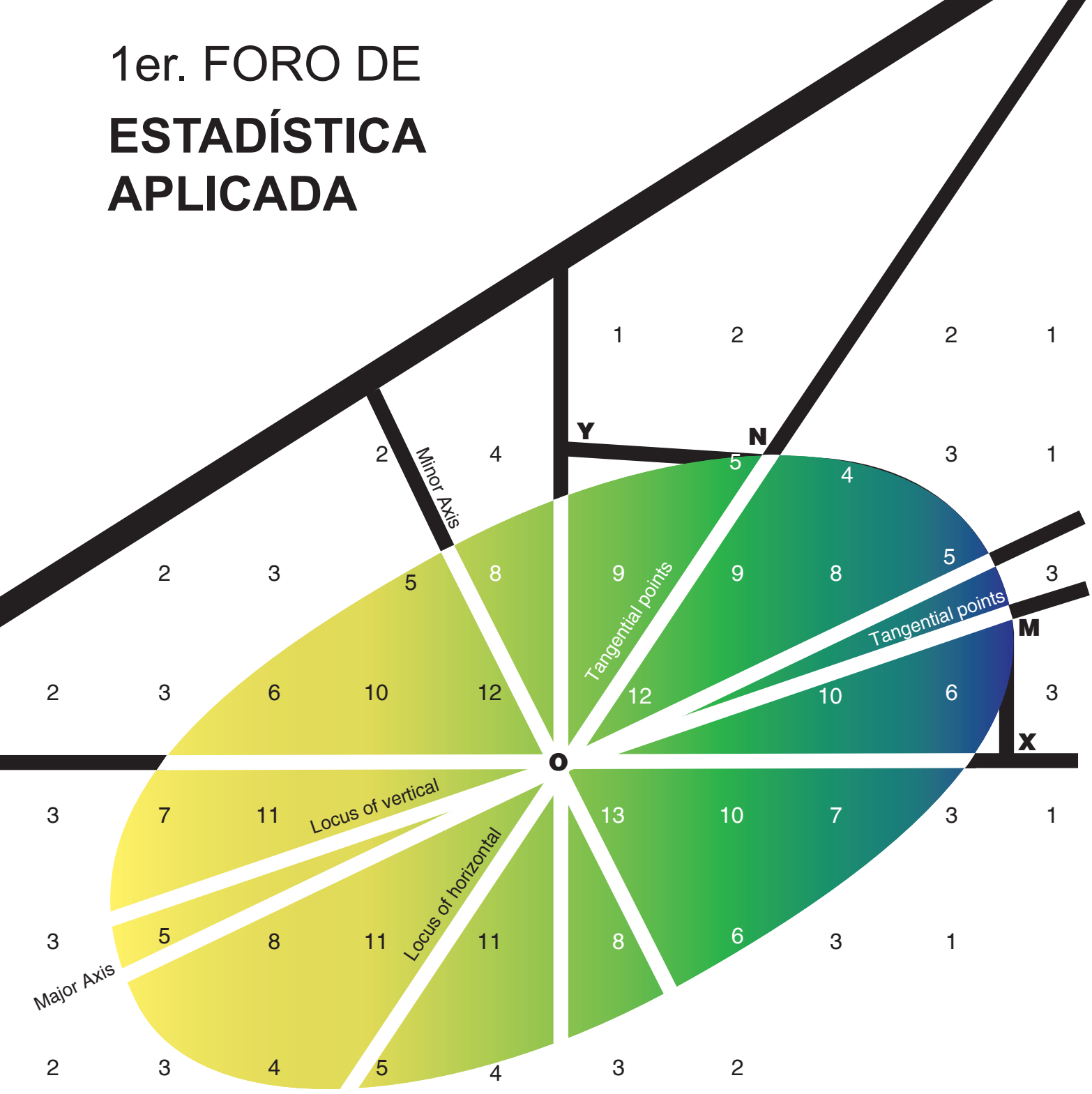
1er. FORO DE
ESTADÍSTICA APLICADA



Unidad Académica de los Ciclos Profesional y de Posgrado
Colegio de Ciencias y Humanidades

1er. FORO DE
ESTADÍSTICA APLICADA

Unidad Académica de los Ciclos Profesional y de Posgrado
Colegio de Ciencias y Humanidades

PRIMER FORO DEL PROYECTO ACADEMICO

ESPECIALIZACION EN ESTADISTICA APLICADA

UNIDAD ACADEMICA DE LOS CICLOS PROFESIONAL Y

DE POSGRADO DEL COLEGIO DE

CIENCIAS Y HUMANIDADES

Septiembre, 1986

I N D I C E

Pág.

PRESENTACION

Durante los días 24, 25, y 26 de Septiembre de 1986 se llevó a cabo en la Unidad de Seminarios Dr. Ignacio Chavez de la U.N.A.M., el Primer Foro de Estadística Aplicada. Este evento fue organizado por la Coordinación de la Especialización en Estadística Aplicada (E.E.A.) de la U.A.C.P. y P. del C.C.H., y contó con el apoyo de la Facultad de Ciencias, del Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas así como del PADEP.

Este evento permitió a estudiantes y egresados del Proyecto de E.E.A, conocer el tipo de trabajos que actualmente se realizan en México, utilizando la estadística como herramienta metodológica.

Asimismo, los participantes provenientes de otras instituciones pudieron apreciar el nivel de los trabajos desarrollados en áreas tales como la Biología, Economía, Demografía, Antropología, Medicina, Ecología y Química.

El foro permitió también un intercambio de experiencias y conocimientos de investigadores que teniendo dominio común de estudios pertenecen a instituciones diferentes.

Numéricamente el evento puede resumirse así: se presentaron un total de 32 ponencias y una mesa redonda. Estuvieron representadas 18 instituciones tanto de educación superior y de investigación como de los sectores público y privado. Cabe resaltar el hecho de que participaron como expositores 18 egresados de diversos programas académicos de la U.A.C.P y P. del C.C.H. De los trabajos expuestos, 15 forman parte de tesis de los niveles de licenciatura, maestría, especialización y doctorado.

El balance de este Primer foro de Estadística Aplicada es alentador. En efecto, los objetivos planteados fueron alcanzados y con esta experiencia estamos ya planeando la organización del Segundo Foro de Estadística Aplicada que se llevará a cabo en el mes de Octubre del

presente año.

En esta memoria se incluyen algunos de los trabajos presentados y que fueron seleccionados en función de su originalidad y del tipo de técnica utilizada.

Se pretende cubrir el espectro de las áreas representadas en el foro y en cuanto a las técnicas, mostrar aplicaciones desde técnicas descriptivas hasta algunas muy particulares.

Con este volumen se espera poder dar un panorama del tipo de estadística que actualmente se emplea en la práctica tanto en el área de investigación como en el de la función pública y privada.

DR. RUBEN HERNANDEZ CID
Coordinador del Proyecto
Especialización en Estadística
Aplicada.

PROGRAMA DEL PRIMER FORO DE ESTADISTICA APLICADA

Miércoles 24

9:30 - 9:45	Registro de Participantes	
9:45	Inauguración por el Lic. Manuel **MARQUEZ** , Director de la U.A.C.P. y P. del C.C.H.	
	NOMBRE (S)	TITULO DE LA PONENCIA
10:00 - 10:20	**Dr. Edmundo BERUMEN** Director General de Estadística Instituto Nacional de Estadística Geografía e Informática (INEGI)	"Encuesta Nacional Ingreso-Gasto de Hogares 1983"
10:25 - 10:45	**Act. Elsa RESANO** INEGI	"Diseño y Construcción de la muestra para la encuesta nacional de empleo urbano 1985"
10:50 - 11:10	**Act. Arturo BLANCAS** INEGI	"Los censos económicos 1986"
RECESO		
11:30 - 11:50	**Act. Alejandro GONZALEZ** Egresado de la Maestría en Ciencias del Mar	"Estructura comunitaria de la fauna decápoda en la plataforma continental del noreste del Golfo de México"
11:55 - 12:15	**Dr. Gustavo VALENCIA** Laboratorio de Estadística Depto. de Matemáticas Facultad de Ciencias.	"Una aplicación del Análisis Discriminante sobre la madurez sexual en truchas arco-iris".

12:20 - 12:40	**Biól. Lucina HERNANDEZ** Instituto de Ecología	"Dimorfismo sexual en la tortuga de Mapimí"
12:45 - 13:05	**Act. José MENDOZA** Laboratorio de Estadística Depto. de Matemáticas Prof. Facultad de Ciencias	"Una aplicación del Análisis Discriminante en clasificación de encinos"
16:00 - 16:20	**M. en C. Guillermo BAZ y DR. Alberto YSUNZA** Depto. de Estadística I.I.M.A.S. Instituto Nacional de Nutrición	"Comparación de patrones de consumo de alimentos : una representación gráfica"
16:25 - 16:45	**DR. Enrique de ALBA** Jefe del Depto. de Matemáticas Instituto Tecnológico Autónomo de México	"Un método bayesiano empírico para detectar cambios en el riesgo de una cartera de inversión"
16:50 - 17:10	**M. en C. Raúl RUEDA** Jefe del area de Estadística Depto. de Matemáticas Universidad Autónoma Metropolitana-Iztapalapa	"Una aplicación de modelo Log-lineal en un estudio comparativo de seis métodos para la determinación de la susceptibilidad a los antimicrobianos en gérmenes anaeróbicos"
17:30 - 17:50	**Act. Jorge ROSASLANDA** Depto. de Matemáticas Colegio de Ciencias y Humanidades SUR	"Galton y Pearson: dos pioneros de la estadística aplicada"
17:55 - 18:15	**Dr. Carlos ALVAREZ** Grupo de filosofía de la Ciencia Depto. de Matemáticas Facultad de Ciencias	"Pensamiento Matemático y Medida"

Jueves 25

9:35 - 9:55	**M. en C. Guillermina ESLAVA** Depto. de Estadística I.I.M.A.S.	"Una aplicación del Escalamiento Multidimensional en un estudio en perinatología"
10:00 - 10:20	**M. en C. Martha de GARAY** Dirección Gral. de Información y Evaluación Secretaría de Salud	"Análisis de la frecuencia de los factores que propician el abandono de la lactancia a seno materno"
10:25 - 10:45	**M. en C. Belem TREJO** Depto. de Estadística I.I.M.A.S.	"Una introducción al análisis de supervivencia"
10:50 - 11:10	**M. en C. Rebeca PONCE DE LEON** Depto. de Estadística I.I.M.A.S. Profesora en la Especialización en Estadística Aplicada	"Niveles y tendencias de la mortalidad infantil en México"
RECESO		
11:30 - 11:50	**Dr. Sergio CAMPOSORTEGA** Subdirector de Información y Pronósticos Consejo Nacional de Población Prof. de la Maestría en Urbanismo Fac. de Arquitectura	"La utilización de la Estadística en Demografía. El ejemplo de la Construcción de tablas Modelo de mortalidad"

11:55 - 12:15	**Act. Mario CORTINA** Instituto de Investigaciones Antropológicas.	"Ajuste de cursos individuales de crecimiento humano"
12:20 - 12:40	**Act. Eduardo CASTAÑO** Depto. de Estadística I.I.M.A.S.	"Una aplicación del Análisis de Correspondencia en evaluación sensorial de alimentos"
12:45 - 13:05	**Dr. Adip SABAG** Centro de Informática Facultad de Contaduría	"Análisis de Discurso Político"
16:00 - 16:20	**Mat. Delfino VARGAS** Prof. de la Especialización en Estadística Aplicada Depto. de Biometría Instituto Nacional de Investigaciones Agrícolas	"Aplicación del Análisis de Correspondencias en el estudio de la interacción medio ambiente-vegetación en el valle de Apatzingán"
16:25 - 16:45	**M. en C. Julio ALFONSO** Depto. de Sistemas Area de Estadística Universidad Autónoma Metropolitana - Azcapotzalco	"Tamaño óptimo de la muestra (en muestreo estratificado) vía programación estocástica"
16:50 - 17:10	**Dra. Thalía HARMONY** Jefa del Programa de Neurociencia ENEP Iztacala	"Neurometría"

Viernes 26

10:00 - 10:20	**M. en C. Jorge DOMINGUEZ** Centro de Investigaciones en Matemáticas Guanajuato, Gto.	"Un paquete conputacional para ajuste de modelos no lineales"
10:25 - 10:45	**Act. Hortensia GONGORA** Depto. de Estadística I.I.M.A.S.	"Estimación de parámetros de modelos ARMA (p,q) usando mínimos cuadrados no-lineales"
10:50 - 11:10	**Act. Estela PATIÑO y Lic. en Ec. David ALERHAND** Egresadosdel Instituto Tecnológico Autónomo de México	"El ciclo económico: una estimación para México, 1960 - 1984.
RECESO		
11:30 - 11:50	**Dr. Jaime JIMENEZ** Modelos Matemáticos de los Sistemas Sociales I.I.M.A.S.	"Estudio comparativo del potencial científico y Tecnológico en México y Hungría"
11:55 - 12:15	**Dra. Teresa LOPEZ** Telmex I.I.M.A.S.	"Modelado de Tráfico Telefónico"
12:20 - 12:40	**Dr. Jaime CURTS** Coordinación del Posgrado ENEP Iztacala Depto. de Estadística	"El Análisis de Residuos en análisis exploratorios de datos"
12:45 - 13:05	**Dr. Ignacio MENDEZ** Rector Universidad Autónoma Chapingo	"La estadística en Agronomía"

16:00 - 18:00	Mesa redonda **Moderador: Dr. Rubén HERNANDEZ** Coordinador de la Especialización en Estadística Aplicada.	"La Estadística en México: sus problemas y sus perspectivas".
18:00	Clausura.	

EL USO DE LA ESTADISTICA EN LA INVESTIGACION DEMOGRAFICA: EL EJEMPLO DE LA CONSTRUCCION DE TABLAS MODELO DE MORTALIDAD.

SERGIO CAMPOSORTEGA CRUZ
CONSEJO NACIONAL DE POBLACION

Introducción

La demografía y la estadística son disciplinas estrechamente relacionadas. Una y otra han contribuido enormemente a su mutuo desarrollo. - La demografía, por su parte, ha impulsado el desarrollo y perfeccionamiento de diversas técnicas estadísticas al plantearle problemas específicos en el estudio de las características demográficas, en tanto que la estadística ha enriquecido el acervo de los métodos y técnicas que - el demógrafo utiliza en el estudio de la población.

En términos generales, la estadística se utiliza en las tres esferas de la investigación demográfica: Generación de Información, Análisis Demográfico e Investigación Causal. La Generación de Información, como su nombre lo indica, corresponde a la recolección de datos demográficos a través de las fuentes clásicas: censos, encuestas y estadísticas vitales. El Análisis Demográfico, por su parte, se identifica con el tratamiento de los datos observados a fin de obtener parámetros útiles de medición, como tablas de mortalidad, tasas específicas de fecundidad, tasas de migración y la elaboración de proyecciones de población y las que se derivan, como es el caso de la proyección de la Población Económicamente Activa, de hogares y de educación, entre otras. Finalmente, la Investigación Causal, que en sí constituye el fin último de la investigación demográfica, trata de examinar las causas de los fenómenos.

En la Generación de Información, la estadística no sólo proporciona las bases teóricas para el diseño de encuestas, sino también métodos efectivos de verificación y control de calidad de los datos demográficos.

En el Análisis Demográfico, la estadística es ampliamente utilizada tanto en las tareas de evaluación, ajuste y corrección de información, como en las de estimación de los parámetros demográficos. En las primeras son frecuentemente utilizadas las técnicas de regresión y los métodos del análisis exploratorio de datos, en tanto que en la etapa de estimación se utilizan los métodos de regresión y algunos modelos estadísticos como el análisis de factores y el de componentes principales.

La Investigación Causal, por su parte, utiliza tanto las diversas técnicas estadísticas de verificación de hipótesis, como los métodos de regresión univariada y multivariada y los análisis de factores, componentes y discriminantes, entre otros.

En las páginas que siguen se presenta un ejemplo de aplicación de la estadística en la demografía: la construcción de las tablas tipo de mortalidad.

1. GENERALIDADES SOBRE LAS TABLAS - MODELO

Dentro del estudio de la mortalidad, una herramienta de gran utilidad es la tabla de vida o de mortalidad, que describe la extinción por muerte de una cohorte de recién nacidos[1/], mediante diversas funciones como las probabilidades de muerte, los sobrevivientes las defunciones, los años vividos por la cohorte y la esperanza de vida.

El comportamiento de la mortalidad por edad ha sido de gran interés tanto para los demógrafos como para los actuarios desde la construcción de las tablas de Graunt y de Halley en la segunda mitad del siglo diecisiete. A partir de entonces, los estudios del fenómeno han tratado de desarrollar una función matemática que describa la experiencia de la mortalidad en toda la vida, enfoque que, sin embargo, no ha sido muy exitoso.

Como alternativa, los demógrafos han ensayado recientemente un nuevo enfoque: la construcción de tablas-tipo de mortalidad, donde en lugar de relacionar los riesgos de defunción solamente por edad, se ha relacionado aquéllos de cierta edad a los cocientes de otras edades o bien a los riesgos observados dentro de otras poblaciones

---

1/ Véase, por ejemplo, Leguina, J., Fundamentos de Demografía, Siglo XXI, Madrid, 1976.

a la misma edad[2/]. De esta forma, las tablas modelo de mortalidad constituyen una respuesta a los repetidos fracasos por encontrar - una función matemática sencilla que describa la evolución por edad de la mortalidad. En términos generales, la idea de las tablas modelo es resumir en una serie de estructuras tipo un amplio conjunto de datos observados.

El desarrollo de las estructuras tipo se ha efectuado mediante diversos procedimientos estadísticos que ilustran la evolución de la demografía y la estadística y constituye un valioso ejemplo de cómo la estadística ha sido utilizada por los demógrafos.

2. ANTIGUAS TABLAS-TIPO DE LAS NACIONES UNIDAS[3/]

Las primeras tablas-tipo se construyeron en la década de los cincuenta por la División de Población de las Naciones Unidas bajo la dirección de U. G. Valaoras[4/]. Estas tablas-tipo se basan en 158

2/ Ver Naciones Unidas. Técnicas indirectas para estimaciones demográficas, Manual X, Departamento de Economía Internacional y Asuntos Sociales, New York, 1983, p. 12.

3/ Naciones Unidas, Schémas de variation de la mortalité selon l'âge et le sexe. Tables-types de mortalite pour les pays sous-developpés, New York, 1956.

4/ Anteriormente, la Liga de las Naciones Unidas efectúa la primera - tentativa de resumir los patrones de mortalidad por edad, vinculando los cocientes, q(x), a la esperanza de vida a los 10 años, e(10), por regresiones lineales. Ver: Notestein, F., et. al., The Future Population of Europe and the Soviet Union. League of Nations, Genéve, 1944.

tablas de mortalidad del período 1891-1950, provenientes principalmente de países desarrollados[5/]. Sin embargo, no todas las tablas tienen la misma calidad de información, e incluso en algunos casos los datos muy deficientes[6/]. Además, algunas tablas han sido suavizadas por diferentes procedimientos[7/].

La construcción de las tablas-tipo asume que el cociente q(x,n) es una función cuadrática del cociente precedente q(x-n, n):

$$q(x,n) = a + (b * q(x-n,n)) + (c * q(x-n) \wedge 2)$$

La estimación de los coeficientes a, b y c se realizó mediante regresión en cadena de los datos de los dos sexos combinados, conforme a las diferencias observadas entre las esperanzas de vida de los hombres y de las mujeres.

5/ 95 tablas pertenecen a Europa, 21 a Asia, 17 a América del Norte, 11 a América del Sur, 8 a Oceanía y 6 a Africa.

6/ Por ejemplo aquéllas de El Salvador 1949-1951, de México 1930 y 1940, de Colombia 1939-1941 y de la India 1891-1901 y 1901-1911.

7/ Ver Clairin, R., et al., La Mortalité dens les pays en développement, Tomo III, Organisation de Coopération et Développement Economiques, Paris, 1980, p. 17.

A propósito de estas tablas podemos formular las observaciones siguientes:

a) El universo de los datos de base reflejan esencialmente la experiencia de las poblaciones europeas o de origen europeo.

b) La utilización de tablas deficientes y de algunas otras - que han sufrido diferentes tipos de ajuste pueden sesgar - los resultados.

c) El tratamiento estadístico de los datos (regresión en cadena) tiene desventajas inherentes, especialmente cuando - -como es el caso- la distribución de los errores no tiene una media igual a cero[8/].

d) Las tablas suponen que hay un sólo esquema de mortalidad, se trata pues de un sistema de un sólo parámetro.

e) Las tablas de cada sexo provienen de las tablas de sexos reunidos, fijándose así la estructura de la mortalidad por sexo y por tanto las diferencias que existen entre ellos.

3. TABLAS-TIPO DE GABRIEL Y RONEN[8/]

Dos años después de la aparición de las tablas de las Naciones Unidas, R. Gabriel e I. Ronen publicaron un artículo donde mostraron que la esperanza de vida estimada a partir de estas tablas se encuentra sobrevaluada en promedio 2,133 años.

La objeción principal a la metodología seguida por Naciones Unidas es la utilización de la regresión en cadena, que tiene la desventaja de acumular los errores $\varepsilon(0)$, $\varepsilon(1)$, $\varepsilon(5)$, $\varepsilon(10)$,..., $\varepsilon(x)$ dentro de la estimación de q(x,n), ya que este cociente se basa, a la vez, en una estimación de q(x-n,n) y así sucesivamente[9/].

Una vez que los autores señalaron la causa del sesgo, calculan, utilizando casi los mismos datos que anteriormente, las mejores estimaciones lineales de los cocientes[10/] a partir del procedimiento tradicional de mínimos cuadrados entre estos cocientes (q(x,n)) -

---

8/ Ver Gabriel, K. et Ronen, I., "Estimates of mortality from infant mortality rates", Population Studies, Vol. 12, No. 2, Londres, 1958.

9/ Las relaciones de las q(x,n) son:
$q(x,n) = a(x) + b(x) * q(x-n,n) + c(x) * (q(x-n,n) \wedge 2) + \varepsilon(x-n)$
$q(x-n,n) + a(x-n) + b(x-n) * q(x-2n,n) + c(x-n) * (q(x-2n,n) \wedge 2) + \varepsilon(x-2n)$
las cuales, evidentemente acumulan errores.

10/ Las mejores estimaciones lineales son aquéllas que no tienen sesgo y minimizan la varianza.

y el cociente de mortalidad infantil q(0,1)):

$$q(x,n) = a(x) + b(x) * q(0,1)$$

Junto con las estimaciones se presenta la correlación y la varianza de las relaciones, que es un gran avance dentro de la construcción de tablas-tipo, ya que permite fijar, con cierto nivel de confianza, los intervalos dentro de los cuales se encuentran los verdaderos valores de los cocientes. A este respecto, se comprueba desgraciadamente que la varianza explicada por las regresiones no es muy grande: después de los 5 años, oscila entre el 62 y el 69 por ciento.

Los modelos de Gabriel y Ronen proporcionan también una estimación alternativa de la esperanza de vida obtenida por regresión entre q(1,0) y e(0)[11/], la cual no coincide con la cifra que resulta de las estimaciones de los cocientes. Este hecho nos muestra que las mejores estimaciones de ciertas funciones de la tabla no conducen a estimaciones insesgadas y de varianza mínima de otros parámetros.

11/ La estimación es:

$e(0) = 75{,}230 - 238{,}08 * q(1,0) + 239{,}46 * (q(1,0) \wedge 2)$

4. TABLAS-TIPO DE COALE Y DEMENY[12/]

Los modelos regionales de la Universidad de Princeton, publicados en 1966, provienen de 192 tablas de mortalidad masculina y femenina, que comprenden el período de 1851 a 1959 (39 de ellas pertenecen al siglo diecinueve y 69 a la posguerra). Estas tablas fueron construidas a partir de las estadísticas vitales y de los censos, en su gran mayoría son nacionales y representan, casi totalmente, la experiencia de las poblaciones de origen europeo (92%). Las 192 tablas han sido seleccionadas de 326 tablas originales entre las cuales se eliminaron aquéllas de dudosa calidad.

Antes de construir las tablas-tipo, los autores establecieron un modelo preliminar a fin de distinguir las distintas familias. El modelo consiste en ordenar de mayor a menor los cocientes de mortalidad de cada edad de las 326 tablas, de modo que las nuevas tablas están constituidas por los cocientes del mismo rango. El examen de las desviaciones de cada tabla y su modelo $(q(x,n) - q(x,n)\ mod)$ produce cuatro tipos de comportamientos:

a) Familia Este. Estas tablas muestran desviaciones del modelo preliminar que se caracterizan por tasas elevadas de mortalidad infantil, así como de altas y crecientes tasas después de los 50 años.

12/ Coale, A. y Demeny, P., Regional model life tables and stable populations, Princeton University Press, New Jersey, 1966.

b) Familia Norte. Esta familia se aparta del modelo preliminar en la medida en que tiene las más bajas tasas de mortalidad infantil y tasas superiores entre los 5 y 40 años. Las desviaciones entre 5 y 40 años se explican por la incidencia de la tuberculosis dentro de las tablas base.

c) Familia Sur. El patrón de mortalidad subyacente se caracteriza por la presencia de elevados niveles antes de los 5 años, tasas más bajas entre 40 y 60 años y más altas - arriba de los 65 años.

d) Familia Oeste. Esta familia, basada en las tablas residuales, no muestra desviaciones sistemáticas del modelo - preliminar.

Identificadas las regiones de mortalidad se calcularon las regresiones lineales entre q(x,n) y log(q(x,n)) y e(10) dentro de las tablas de cada familia:

$$q(x,n) = A(x) + B(x) * e(10)$$

$$\log (10000 * q(x,n)) = AA(x) + BB(x) * e(10)$$

Los valores de las q(x,n) provenientes de las regresiones logarítmicas resultaron siempre superiores a las regresiones sin transformaciones en los extremos de la esperanza de vida y a la inversa en las esperanzas intermedias. Los autores decidieron retener las - q(x,n) de la regresión simple antes de la primera intersección, -

las q(x,n) de los logaritmos después de la segunda intersección y la media de las dos en el centro; dado que la regresión logarítmica está más cerca de los datos dentro de las esperanzas de vida elevadas y, en cambio, dentro de las esperanzas más bajas, aquella que ajusta mejor es la regresión lineal.

Las tablas finalmente utilizadas en las regresiones fueron 31 en el modelo Este, 9 en el Norte, 22 en el Sur y 130 en el Oeste. En el primer caso, las esperanzas de vida van de 36.6 (Baviera, 1878) a - 72.3 años (Checoslovaquia, 1958); en el segundo de 44.4 (Suecia, - 1851-1860) a 74.7 años (Noruega 1951-1955); en el modelo Sur de - 35.7 (España, 1900) a 68.8 años (Italia del Sur, 1954-1957), y por último, en el Oeste de 38.6 (Taiwan, 1921) a 75.2 años (Suecia, - 1959)[13/].

El enfoque de Coale y Demeny permite librar ciertas críticas propias de las tablas precedentes, sin embargo diversas advertencias pueden ser formuladas:

a) El universo de tablas que sirve de base a los modelos reflejan esencialmente la experiencia de las poblaciones de origen europeo (92%), en consecuencia las cuatro familias

---

13/ Nations Unies, Indirect techniques for demographic estimations, - cp. cit.

no cubren enteramente la diversidad de situaciones posibles[14/].

b) "El sistema permanece todavía poco flexible, ya que las regresiones se basan todas sobre una entrada única (e(10)) - en el seno de cada familia"[15/].

c) El tamaño de la muestra dentro de las familias Norte y Sur no permiten generalizaciones muy satisfactorias.

d) La utilización de entradas diferentes a e(10) produce ligeros errores de estimación[16/].

5. TABLAS-TIPO DE S. LEDERMANN[17/]

Ledermann y Brass en 1959[18/] y Burgeois-Pichat un poco más tarde[19/] identificaron, a partir del análisis de factores, aquéllos que más explican la variación de la mortalidad dentro de diferentes tablas.

14/ Ver por ejemplo:
Adlakha, A., "Model life tables: an empirical test of their applicability to less developed countries", Demography, Vol. 9, No. 4, 1972.

15/ Clairin, R., et al., op. cit., p. 19.

16/ Le Brass, H., "Avant-propos" in Lederman, S. Nouvelles tables-types de mortalité, P. U. F., Paris, 1969.

17/ Ledermann, S., Nouvelles tables-types de mortalité, P.U.F., Paris, 1969.

18/ Ledermann, S. et Brass, J., "Les dimensions de la mortalité", Population, Vol. 14, No. 4, Paris, 1959.

19/ Burgeois-Pichat, J., "Factor analysis of sex-age specific death rates", Population Bulletin of the United Nations, No. 6, New York, - 1962.

El primero y más importante factor está asociado al nivel general de la mortalidad, el segundo se refiere a la relación entre la mortalidad infantil y adulta, el tercero corresponde a la mortalidad de los ancianos, el cuarto se refiere al patrón de la mortalidad debajo de los 5 años y, por último, el quinto factor está asociado a las diferencias entre la mortalidad masculina y femenina dentro de las edades que van de 5 a 70 años[20/].

A partir de los resultados de este análisis y mediante la utilización del análisis de regresión de 154 tablas de mortalidad, S. Ledermann construyó en 1969 nuevas tablas tipo de mortalidad con 1 y 2 parámetros de entrada.

El conjunto de las 154 tablas de base es muy similar al que utilizó las Naciones Unidas en 1956, si bien se suprimen algunas de las tablas de insuficiente calidad.

Las tablas tipo se construyeron a partir de los cocientes de mortalidad, estimados mediante eucaciones de regresión de la forma:

$$\ln (q(x,n)) = A(x) + B(x) * \ln(E)$$

20/ Cf. Naciones Unidas, Indirect techniques for demographic estimations, op. cit., p. 16.

para las tablas de una entrada, y

$$\ln(q(x,n)) = AA(x) + BB(x) * \ln(E1) + CC(x) * \ln(E2)$$

para las tablas de dos entradas. Dentro de estas ecuaciones, E, - E1 y E2 representan las variables independientes (entradas) y A(x), B(x), AA(x), BB(x) y CC(x) los coeficientes estimados por las regresiones.

Suponiendo normalidad en la distribución de los logaritmos de los cocientes de mortalidad, Ledermann presenta además de los valores - de los cocientes (mediana o media geométrica), la extensión de la zona de dispersión, dentro de la cual se encuentra el 95% de las - observaciones (+ - 25, ó 5 es la desviación estándar)[21/]. Esta medida, evidentemente, sólo se refiere a los conjuntos de las tablas utilizadas dentro de las regresiones y por tal motivo, no cubre todas las situaciones posibles.

El autor efectúa, además un excelente análisis estadístico de los modelos utilizados en la construcción de las tablas-tipo. De esta manera muestra que: a. La utilización de una entrada diferente a

---

21/ A este respecto, el autor menciona que la dispersión es medida sólo si es aproximativamente la misma para todas las observaciones y si es simétrica; esto es, de acuerdo a la utilización de los logaritmos de los cocientes, pues la media geométrica coincide con la mediana y está muy próxima a la moda, de esta forma dado la distribución - gausiana de los cocientes logarítmicos, el 95% de las distribuciones se encuentran dentro de la zona que va de -25 a +25. Es importante señalar que dentro de las tablas de las Naciones Unidas, Gabriel y Ronen y Coale y Demeny, estas condiciones no son satisfechas.

la indicada causa sesgo; b. Aunque los valores centrales de los cocientes son estimados sin sesgo, los sobrevivientes que se pueden deducir son estimaciones sesgadas, y a la inversa. En efecto no es posible obtener estimaciones insesgadas de las dos series - cuando no son combinaciones lineales los cocientes y los sobrevivientes[22/]; c. En ciertos valores de las entradas, el modelo puede proporcionar sesgos, en la medida que no se adapta suficientemente a la distribución de los puntos[23/]; y d. La esperanza de vida al nacimiento que resulta de los cocientes estimados no coincide con el valor de la entrada[24/].

Los resultados comprenden 7 tablas de una entrada (e_0, q(0,1), - q(0,5), q(0,15), q(30,20), q(45,20) y m(50 +) y 3 de dos entradas (q(0,5) y q(45,20), q(0,15) y q(30,20), q(0,15) y q(50 +)) las - cuales, a excepción del cociente q(45,20), se refieren a datos de sexos combinados.

El minucioso trabajo estadístico aplicado a las tablas de base, - así como la extensión de las entradas y la introducción de otra en

---

22/ Ledermann, S., op. cit.
Este hecho se produce porque la media de una serie de relaciones no es igual, en general, a las relaciones de la suma de los numeradores y de los denominadores.

23/ Es el caso de las estimaciones de q(0,1), q(1,4) y q(5,5) calculados por Gabriel y Ronen. En efecto, para los valores de las entradas - débiles, dan cocientes negativos.

24/ Esto debe ser, como dentro del punto b, a la no linearidad de las - relaciones entre los cocientes y los sobrevivientes a la esperanza de vida de cierre de una tabla. Esta esperanza "es una estimación sesgada de la esperanza de vida central de todas las situaciones - que tienen en común el valor de entrada considerado", Ledermann, - S., op. cit.

trada cuantitativa permiten escapar a ciertas críticas emitidas a propósito de las tablas-tipo precedentes, sin embargo se pueden señalar diversos puntos discutibles:

a) Las tablas de base recuperan esencialmente la mortalidad europea y algunas de ellas no tienen una calidad aceptable.

b) Las entradas seleccionadas no son muy adecuadas, pues en general no corresponden a las necesidades del análisis demográfico actual[25/].

c) La utilización de datos de sexos combinados dentro de la casi totalidad de las entradas fijan la estructura de la mortalidad por sexo, los cuales no son siempre satisfactorias. Así, por ejemplo, es casi imposible obtener, a partir de estas tablas, una esperanza de vida masculina superior a la femenina[26/].

6. NUEVAS TABLAS-TIPO DE LAS NACIONES UNIDAS[27/]

Estas nuevas tablas-tipo fueron publicadas por las Naciones Unidas en 1982. Los datos de base parten de un banco de datos sobre mor-

---

25/ Cf. Clairin, R., et al., op. cit., p. 20.

26/ Cf. Page, H. et Wunsch, G., "Parental survival data: some results of the application of Ledermann's model life tables", Population Studies, Vol. 30, No. 1, Londres, 1976.

27/ Nations Unies, Model life tables..., op. cit.

talidad para países subdesarrollados establecido por la Organización de Cooperación y Desarrollo Económico.

En vista que los datos demográficos de los países subdesarrollados son, en general, de una calidad muy pobre, se llevaron a cabo minuciosos trabajos de evaluación, ajuste y corrección con el fin de crear un conjunto de tablas libres de errores. Según los autores, la filosofía subyacente a la construcción de las tablas-tipo es que los modelos pueden solamente ser confiables si el conjunto de las tablas de base lo es[28/]

Los procedimientos de evaluación utilizados son dos tipos: pruebas de coherencia interna y comparación con otras fuentes. Entre los primeros, se examinaron las distribuciones por edad y sexo, la relación de masculinidad, los esquemas de mortalidad, las disminuciones de las tasas de mortalidad más allá de los 50 años relacionándolos a las funciones de Makenham y Gompertz y a los resultados de los métodos de cobertura (Brass, etc.). En lo que concierne a las comparaciones, se utilizaron las encuestas retrospectivas, los "matching survey" y los resultados de los métodos indirectos. A partir de estos análisis, los autores construyeron 36 tablas de mortalidad por cada sexo (72 en total), las cuales comprenden esperanzas de vida entre 37 y 76 años. Del total de tablas, 19 pertenecen a 11 países asiáticos, 16 a 10 países de América Latina y 1 a Africa.

28/ Ibid, p. 2 (tr. SCC).

Las tablas se repartieron en 4 familias diferentes mediante tres procedimientos estadísticos y uno gráfico ($R(x) = q(x,n)/qw(x,n)$; donde $qw(x,n)$ es el cociente de la familia Oeste de las tablas de Princeton y $q(x,n)$ el cociente de la tabla respectiva. El patrón latinoamericano, compuesto por 9 tablas de esta región y 6 de países de Asia del Sur, se caracteriza con relación a la familia Oeste, por una fuerte mortalidad relativa en las primeras edades y entre 15 y 45 años y por un débil nivel de las edades elevadas[29/]. El patrón chileno, compuesto por 3 tablas de este país, presenta altas tasas de mortalidad infantil, pero niveles bajos en los niños mayores de un año. El esquema del sudeste asiático, formado por 4 tablas, muestra una muy fuerte mortalidad relativa antes de los 15 años y en la tercera edad[30/]. El patrón del Extremo Oriente, que se caracteriza por una muy fuerte mortalidad en las edades elevadas, está compuesto, en el caso masculino, por 4 tablas de Trinidad y Tobago y Guayana y 5 de Hong Kong, Corea y Singapur y en las mujeres, por una tabla de Singapur y 4 de los países de América ya citados. Finalmente, la mortalidd del patrón general es cercano a la familia Oeste de las tablas de Princeton[31/].

29/ Este comportamiento está dado por el exceso de muertes infantiles causadas por la diarrea y los parásitos; la importancia de los accidentes y de la baja incidencia de enfermedades cardiovasculares. Ver Nations Unies, Indirect techniques for demographic estimation, op. cit, pp. 19-21.

30/ Este patrón, que es muy próximo de la familia Sur de Princeton, parece estar en relación con un gran número de muertes infantiles atribuibles a infecciones, parásitos y diarrea; y de ancianos causadas por la diarrea y las enfermedades respiratorias. Ver Ibid, p. 12.

31/ Excepción hecha de los niveles de mortalidad bajos, cuyas diferencias pueden explicarse por los riesgos de las extrapolaciones.

La técnica de construcción de las tablas-tipo recurre al modelo de componentes principales. De esta forma, al interior de cada región, se ajustaron ecuaciones de la forma:

$$Y(x,n) = Ur\ (0,x) + \sum_{i=1}^{k} a(i) * Ur(i,x)$$

donde Y(x,n) es el logito del cociente observado, es decir Y(x,n) = logit (q(x,n)) = 0,5 * ln (q(x,n)/(1-q(x,n))); Ur(0,x) es el patrón medio de la región r (en términos de logitos); Ur(i,x) representa las desviaciones de los datos observados con relación a la media[32/], y a(i) los coeficientes que indican el peso de estas desviaciones.

Las tablas publicadas corresponden a la aplicación de un componente (k=1), el cual explica el 89% de la variación en las tablas masculinas y el 91% en las femeninas.

32/ Ur(1,x) representa las desviaciones en relación a Ur(0,x); Ur(2,x) las desviaciones en relación a la aplicación del primer componente, etc. En términos demográficos, Ur(0,x) puede ser identificado como el esquema medio de mortalidad: Ur(1,x) como las desviaciones típicas de este esquema a medida que el nivel de mortalidad cambia; y - Ur(2,x) y Ur(3,x) como las desviaciones que no se explican enteramente por las desviaciones de los niveles. Ur(2,x) está asociado - a las diferencias entre la mortalidad de 0 a 4 años y la mortalidad más allá de 5 años.

Al proporcionar las desviaciones promedio U(i,r) (i=1,2,3), la publicación de las Naciones Unidas permite la construcción de nuevos modelos de mortalidad[33/]. En particular, para el caso mexicano hemos construido un nuevo patrón medio de la mortalidad mexicana a partir de las tablas corregidas de 1940, 1950, 1960, 1970 y 1980[34/] y con la utilización de este procedimiento y las desviaciones mencionadas hemos desarrollado un nuevo conjunto de tablas tipo que corresponden a la experiencia nacional.

El enfoque de las Naciones Unidas, si bien altamente sofisticado, conlleva algunas críticas, en particular por el reducido número de tablas que se utilizan. El caso extremo es el patrón chileno que se deriva únicamente de tres observaciones.

33/ Ver Naciones Unidas, Model life tables for developing countries, op. cit., pp. 16-27.

34/ Ver Camposortega,

C O N C L U S I O N E S

El desarrollo de las tablas modelo de mortalidad es un ejemplo muy ilustrativo de la utilización de técnicas estadísticas cada vez más perfeccionadas en los estudios demográficos. En efecto, la búsqueda de estructuras tipo en el comportamiento de la mortalidad por edad sólo ha sido posible mediante el uso de la estadística. Las técnicas aplicadas van desde las imperfectas regresiones en cadena hasta el sofisticado análisis por componentes principales.

La solución estadística a este tipo de problema requiere de un amplio conocimiento de esta disciplina. Su desconocimiento puede ocasionar diversos errores como en los que se ha incurrido en el desarrollo de las tablas-tipo: regresión en cadena, poca representatividad, etc.

En la demografía existen como éste, numerosos problemas que requieren la aplicación de técnicas estadísticas y que para el demógrafo, muchas veces, constituyen retos inalcanzables. Es por ello que los estudios interdisciplinarios donde participen especialistas en estadística son indispensables para un mejor análisis de la problemática poblacional.

UNA APLICACION DEL ANALISIS DE CORRESPONDENCIAS EN EL ANALISIS SENSORIAL DE ALIMENTOS

por

Eduardo Castaño Tostado

IIMAS - UNAM

En el presente trabajo se presentan dos aplicaciones en el àrea de tecnologìa de alimentos,especificamente en el campo de la evaluaciòn sensorial.A continuaciòn se presenta la descripciòn general de la evaluaciòn sensorial de una fruta llamada Jiotilla.

1.Aplicaciòn.

Interesa el estudio de esta fruta considerada exòtica y perteneciente a la familia de las cactàceas, por sus propiedades alimenticias y porque representa una fuente de ingresos adicionales para las comunidades rurales que tienen acceso a ella.Se prepararon 13 tratamientos de esta fruta:

TRATAMIENTO	IDENTIFICADOR
1.En fresco(control)	FFRE
2.Confitada en estufa	CEST
3.Confitada en sol	CSOL
4.Confitada al aire	CAIR
5.Mermelada	MERM
6.Almibar	ALMI
7.Confitada por proceso lento	CPLE
8.Confitada por proceso ràpido	CPRA
9.Glaseada por proceso lento	GPLE
10.Glaseado por proceso ràpido	GPRA
11.Confitada y envasada en cloruro de polietileno	CECP
12.Confitada y envasada en celofàn	CECE
13.Confitada y envasada en cloruro de polivideno.	CECV

Estos tratamientos fueron evaluados por un panel de 15 jueces no entrenados,que calificaron segùn una escala organolèptica(hedònica) de nueve categorias:

CATEGORIA	IDENTIFICADOR
Gusta extremadamente	GEX
Gusta mucho	GMU
Gusta moderadamente	GMO
Gusta ligeramente	GLI
Indiferente	IND
Disgusta ligeramente	DLI
Disgusta moderadamente	DMO
Disgusta mucho	DMU
Disgusta extremadamente	DEX

Esto fue realizado en olor,sabor,color,textura y aspecto de cada uno de los trece tratamientos aplicados a la fruta;se utilizò la tècnica estyadistica denominada Anàlisis de Correspondencias de la que los detalles de su utilizaciòn se dan en el segundo apartado de esta nota.Sòlo son presentados los resultados de las primeras dos evaluaciones en fuciòn del espacio.

1.1 OLOR.

Los datos se muestran en el cuadro 1.1.Se evidencia la nula contabilizaciòn desde DMO hasta DEX por lo que no se consideran en el anàlisis.La representaciòn producida por A.C. se muestra en sus dos primeros ejes en la figura 1.1.

En forma descriptiva se observa que entre estos dos primeros ejes se tiene un 73% de la variaciòn en los datos;respecto a la calidad de cada perfil en la representaciòn con los dos primeros ejes, los perfiles renglòn,es decir,las categorias de la escala de preferencias,en el cuadro 1.2 las contribuciones absolutas y

las correlaciones.Así,en general están bien representados en estos dos ejes,ya que las contribuciones relativas son mayores al 0.5 salvo el caso de DLI.En cuanto a la contribución a los ejes,en el primero de ellos GEX y GLI son los preponderantes sucediendo esto también en el segundo eje,es decir,que ambos ejes son representantes de las posiciones relativas al gusto extremo y al gusto ligero.

En cuanto a los perfiles columna,los tratamientos,se tiene los resultados en cuanto a contribuciones absolutas y correlaciones en el cuadro 1.3.

En vista de las correlaciones,FFRE,CAIR,MERM y ALMI son tratamientos mal representados;los demàs en general no tienen mayores problemas.En cuanto a las contribuciones absolutas,CECP y GPLE son los que dan primordialmente la dirección al eje uno;respecto al segundo eje se unen al grupo anterior CEST y CECE.

En cuanto a la orientación de los trece tratamientos respecto a la escala de preferencias se tiene que GPLE descolla hacia el gusto extremo siguiendo en orden descendente CPLE,GPRA,FFRE,CPRA y CEST.Posteriormente se agrupan hacia un gusto moderado MERM,ALMI,CSOL y CAIR.Por último CECV,CECE y CECP son poco

aceptados.

1.2. Sabor.

Los datos se muestran en el cuadro 1.4.La representación gráfica producida por el A.C. se muestra en la figura 1.2.

Descriptivamente,los dos primeros ejes contemplan el 71% de la variación en los datos;en cuanto a la representación de las categorías de la escala de preferencias se tiene en base al cuadro 1.5,los siguientes comentarios:

Se tiene que GMO y DEX tienen graves problemas en su representación.En cuanto a la contribución absoluta,en el primer eje sobresalen DLI,DMO y GMU con lo que este eje muestra la posición relativa entre el disgusto moderado y el gusta mucho;en el segundo eje nuevamente DMO pesa pero ahora superado por GLI.Respecto a los perfiles columna,es decir los tratamientos a la fruta se tiene en base al cuadro 1.6,lo siguiente:

Se tienen problemas en la representación de FFRE,CEST,CSOL,CAIR y MERM(correlación< 0.5).En cuanto a las contribuciones absolutas CECP y CECV son los de mayor influencia en ambos ejes.en cuanto a la orientación específica de los tratamientos a la fruta respecto a la escala de preferencias,se tiene un agrupamiento fuertemente

aglutinado · alrededor de GEX y GMU,estando constituido por CPLE,GPLE,GPRA y CEST siguiendo MERM,FFRE,ALMI y CSOL.Hacia GLI se tiene a CECE,CPRA,CAIR y CECV.Por último,se separa totalmente CECP hacia DMO y DMU.

2. Anàlisis de Correspondencias,una introducciòn.

El Anàlisis de Correspondencias es resultado de desarrollos geomètricos y algebraicos.Existen distintos enfoques del A.C. en su utilizaciòn;uno de estos enfoques debido a la escuela francesa, y que se sigue en este trabajo,se debe primordialmente a los trabajos de Jean Paul Benzècri.Esta tècnica se sitùa en el anàlisis multivariado de datos y su objetivo principal es la representaciòn gràfica de las relaciones màs relevates que se hayan en forma subyacente en una tabla de contigencia de dos o màs criterios de clasificaciòn.

Es propòsito de esta secciòn el presentar los detalles màs importantes tanto algebraicos como geomètricos del A.C. y para ello se sigue la notaciòn de Greenacre(1984).La presentaciòn por sencillez se harà para el caso de dos criterios de clasificaciòn en la tabla de contigencias,aunque la generalizaciòn al caso multicriterio es inmediata,ya que a fin de cuentas las conclusiones suceptibles de obtenerse mediante el A.C. son sòlo sobre las interacciones de orden uno entre los criterios de clasificaciòn de la tabla bajo estudio(Greenacre,1984).

2.1 Notación.

Sea N una tabla de contingencias con pq celdas y con frecuencias n_{ij} $i=1...,p+1;j=1,...,q+1$ $(p>q)$,y denote por n a

$$n=\sum_i \sum_j n_{ij} \qquad (2.1)$$

Llame P a la matriz de frecuencias relativas

$$P=(1/n)N \qquad (2.2)$$

y a

$$\underline{r}=P\underline{1} \text{ y } \underline{c}=P'\underline{1}$$

donde $\underline{1}$ es un vector de unos con las dimensiones adecuadas al caso,$\underline{r}$ representa el vector de sumas por renglón de P y $\underline{c}$ el vector de sumas por columna de la misma matriz.Denote a los elementos de estos dos vectores como $p_{i.}$,$i=1,...,p+1;p_{.j}$,$j=1,...,q+1$ respectivamente.

Se construyen las siguientes matrices:

$$R=D_r^{-1}P \ , C=D_c^{-1}P' \qquad (2.3)$$

donde

$$D_r=\text{diag}(\underline{r}),D_c=\text{diag}(\underline{c}) \qquad (2.4)$$

Los renglones de la matriz R ,denotados por $\{\underline{r}_i, i=1,...,p+1\}$ son llamados los perfiles renglón de la matriz P;analogamente,los renglones de la matriz $C(\underline{c}_j, j=1,...,q)$ son nombrados perfiles columna.Así,el i-ésimo perfil renglón es un vector cuyas entradas representan la distribución discreta marginal en el i-ésimo

renglòn(i=1,...,p+1)de la matriz P;analogamente en el caso de los perfiles columna.En vista de lo anterior la estructura relativa de la tabla representada en dos formas por estos conjuntos de perfiles es invariante ante el tamaño total de la tabla n.

Ya que los perfiles renglòn como los perfiles columna pueden considerarse como vectores de q+1 y p+1 entradas respectivamente,el objetivo del A.C. es buscar un subespacio de rango menor en los espacios respectivos,en los que las proyeccciones de los corespondientes perfiles representen las asociaciones màs relevantes que se dan en los espacio q+1-variado y p+1-variado respectivamente(figura 2.1).Posteriormente a la obtenciòn de estos dos subespacios el A.C. obtiene una representaciòn gràfica conjunta de ambos subespacios.

Si se denota por $d_m(\underline{v},\underline{w})$ a la distancia de $\underline{v}$ a $\underline{w}$ bajo la mètrica m,un criterio para encontrar el subespacio(S*)en el espacio de perfiles renglòn es

$$\min_{S^*} \sum_{i=1} w_i \, d_m(\underline{r}_i, \underline{s}_i) \quad \underline{s}_i \in S^* \qquad (2.5)$$

donde w_i es un peso diferencial por perfil y $\underline{s}_i$ la proyecciòn ortogonal de $\underline{r}_i$ en S*.

En el A.C. se escoge $w_i = p_i$. y a la mètrica m se le asigna la

matriz D_c;es decir que los perfiles se ponderan en función de su frecuencia relativa total y las q+1 coordenadas de cada uno de estos por su frecuencia relativa acumulada por estas en todos los perfiles. Así, (2.5) queda expresado como

$$\min \|R-S^*\|_{r,c} = \min \sum_i p_{i.} (r_i - s_i)' D_c^{-1} (r_i - s_i) \qquad (2.6)$$

Analogamente para los perfiles columna;es decir,los renglones de la matriz C,se tiene que el criterio utilizado por el A.C. para encontrar el subespacio de mejor ajuste(T^*)es

$$\min \|C-T^*\|_{r,c} = \min \sum_j p_{.j} (\underline{c}_j - \underline{t}_j)' D_r^{-1} (\underline{c}_j - \underline{t}_j) \qquad (2.7)$$

con $\underline{t} \in T^*$.

Los centroides de los perfiles renglòn y de los perfiles columna son respectivamene $\underline{c} = R\underline{r}'$ y $\underline{r} = C'\underline{c}$.Se puede demostrar que èstos estàn contenidos en los planos S^* y T^* por lo que se parte sin pèrdida de generalidad de los perfiles centrados(Greenacre,1984)

$$R-\underline{1}\underline{c}', C-\underline{1}\underline{r}'. \qquad (2.8)$$

2.2 Descomposición en Valor Singular y el A.C.

Suponga que $R-\underline{1}\underline{c}'$ es de rango K ($K<=\min(p,q)$);se pueden encontrar K vectores ortogonales tales que(Green y Carroll,1976):

$$R-\underline{1}\underline{c}' = FB' \qquad (2.9)$$

donde $B=(\underline{b}_1, \underline{b}_2, \ldots \underline{b}_k)$ son los ejes principales de los perfiles

renglòn y F las coordenadas de los perfiles renglòn centrados,respecto a B .Analogamente

$$C-\underline{1}\underline{c}' = GA' \qquad (2.10)$$

con $A=(\underline{a}_1,\underline{a}_2,\ldots,\underline{a}_K)$ ejes principales de los perfiles columna y G las coordenadas respectivas.Ahora,

$$D_r(R-\underline{1}\underline{c}') = P-\underline{r}\underline{c}'$$

$$D_c(C-\underline{1}\underline{r}') = P'-\underline{c}\underline{r}' \qquad (2.11)$$

con lo que encontrar A y B son problemas interrelacionados. La descomposiciòn en valor singular da la pauta para encontrar estos dos conjuntos de vectores.Esta herramienta algebraica provee las matrices $L(p \times k)$, $M(q \times k)$ y $D_u(K \times K)$ tales que

$$P-\underline{r}\underline{c}' = LD_u M' \qquad (2.12)$$

sujeto a que

$$L'D_r^{-1}L = M'D_c^{-1}M = I(K \times K) \qquad (2.13)$$

Las columnas de la matriz L son K vectores ortogonales bajo la mètrica D_r^{-1}, y constituyen una base ortonormal para los renglones de $P-\underline{r}\underline{c}'$, mientras que las columnas de M son K vectores ortogonales bajo la mètrica D_c^{-1}, siendo una base ortonormal para

las columnas de la matriz $P'-\underline{r}\underline{c}'$. Por último, D_u es una matriz diagonal con elementos $(u_1 \geq u_2 \geq \ldots \geq u_K > 0)$ llamados valores singulares. En vista de (2.12)-(2.13) se tiene que

$$P-\underline{r}\underline{c}' = \sum_{k=1}^{K} u_k \underline{m}_k \underline{l}_k' \qquad (2.14)$$

Dado el carácter aditivo en (2.14) se puede aproximar la matriz (2.11) por $K^* < K$ de los vectores de M y L. Si $K^*=2$ se tiene una matriz de rango 2 que aproxima a la matriz de rango K. Esta forma de aproximar a (2.12) es óptima bajo el criterio

$$\min \|A-X\| = \min \sum_i p_{i.} (\underline{a}_i - \underline{x}_i)' D_c (\underline{a}_i - \underline{x}_i) \qquad (2.15)$$

con $A=P-\underline{r}\underline{c}'$ (Green y Carroll, 1976).

De (2.12) y (2.11) se tiene que las columnas de M representan una base para los renglones de la matriz

$$R-\underline{1}\underline{c}';$$

si se denota por F las coordenadas respectivas, es decir,

$$FM = (R-\underline{1}\underline{c}') \qquad (2.16)$$

por (2.13) se tiene que

$$F = (R-\underline{1}\underline{c}') D_c^{-1} M \qquad (2.17)$$

Analogamente, si G denota las coordenadas de los renglones de

$$(C-\underline{1}\underline{r}')$$

respecto a L,

$$G = (C-\underline{1}\underline{r}') D_r^{-1} L . \qquad (2.18)$$

Si K*=2 se tendrà

$$F2=(R-\underline{1}\underline{c}')D_c^{-1}M2 \qquad (2.19)$$

$$G2=(C-\underline{1}\underline{r}')D_r^{-1}L2 \qquad (2.20)$$

serán las coordenadas de los perfiles renglòn y de los perfiles columna en el subespacio de rango dos que mejor aproxima las relaciones en los espacios q y p variados respectivamente,bajo el criterio (2.15),donde M2 y L2 representan los dos primeros ejes principales respectivos.Con estos resultados se tendràn las representaciones gràficas de los perfiles renglòn y de perfiles columna.Debe remarcarse que estas representaciones son en dos espacios diferentes;el objetivo final del A.C. es la representaciòn simultànea de ambos conjuntos,que formalmente no es posible realizar,pero que,a pesar de lo anterior,en la siguiente secciòn se muestran las expresiones que permitiràn establecer tal representaciòn conjunta ad hoc.

2.3 Fòrmulas de Transiciòn.

Las expresiones (2.17)-(2.18) estàn relacionadas mediante las llamadas fòrmulas de transiciòn.Para mostrar lo anterior,primero de (2.17) y (2.11) se tiene que

$$F=D_r^{-1}(P-\underline{r}\underline{c}')D_c^{-1}M \ . \qquad (2.21)$$

Ahora de (2.12),

$$(P - rc')D_c^{-1}M = LD_u$$

con lo que

$$F = D_r^{-1}LD_u . \qquad (2.22)$$

En forma anàloga,

$$G = D_c^{-1}MD_u . \qquad (2.23)$$

De (2.21),(2.23) y (2.12) se tiene que

$$F = RGD_u^{-1} \qquad (2.24)$$

$$G = CFD_u^{-1}$$

llamadas las fórmulas de transiciòn;estas expresiones significan que las coordenadas de un conjunto de perfiles se puede expresar en función de las coordenadas del otro conjunto de perfiles.Para explicar las implicaciones de estas expresiones,considere la coordenada del i-èsimo perfil renglòn respecto al k-èsimo eje principal respectivo denotada por f_{ik};de acuerdo con (2.24),

$$f_{ik} = \sum_{j=1} r_{ij} g_{jk} / u_k \qquad (2.25)$$

donde r_{ij} es la entrada j del perfil i,g_{jk} la coordenada del perfil columna j en su eje principal k y u_k el valor singular respectivo.En vista de (2.25),se tiene que f_{ik} es una combinaciòn de las coordenadas de los perfiles columna con respecto a su eje k.Geomètricamente,un perfil renglòn tenderà a una posiciòn que

corresponderà a las coordenadas de los perfiles columna màs importantes respecto a aquel.Las fòrmulas de transiciòn permiten asi la superposiciòn de gràficas de los dos tipos de perfiles para su interpretaciòn conjunta.

2.4 Interpretaciòn de los Valores Singulares .

La variaciòn estimada de los perfiles renglòn en el eje k es

$$\sum_i p_{i.}(f_{ik}-f_{.k})^2 \qquad k=1,\ldots,K \qquad (2.26)$$

$$f_{.k}=\sum_i p_{i.} f_{ik} \qquad k=1,\ldots,K \qquad (2.27)$$

Puede mostrarse que

$$(f_{.1},f_{.2},\ldots,f_{.k})=\underline{0}$$

con lo que (2.26) es

$$\sum_i pi.f_{ik}^2 \quad . \qquad (2.28)$$

La expresiòn (2.28) puede verse como

$$(f_{1k},f_{2k},\ldots,f_{pk})\begin{bmatrix} p_{1.} & 0 & 0\ldots0 \\ 0 & p_{2.} & 0\ldots0 \\ 0 & 0 & 0\ldots p_{K.} \end{bmatrix}\begin{bmatrix} f_{ik} \\ f_{2k} \\ f_{pk} \end{bmatrix} \qquad (2.29)$$

el producto del primer renglòn de F por sí mismo ponderado por la matriz D_r,es decir,como parte del producto de matrices $F'D_rF$.

De (2.22),la expresiòn (2.28) desarrollandola se llega a

$$F'D_r F=D_u L'\bar{D}_r^{-1} D_r \bar{D}_r^{-1} L D_u =D_u^2.$$

y

$$\text{traza}(F'D_r F)=\text{traza}(D_u^2). \qquad (2.30)$$

El sumando correspondiente a u_k es precisamente (2.28),con lo que u_k^2 es el estimador de la varianza de los perfiles renglòn respecto al eje principal $\underline{m}_k$,k=1,...,K. Si se recuerda los valores singulares estàn ordenados,por lo que en vista de la expresiòn (2.30),$\underline{m}_1$ serà el eje de mayor variaciòn en las coordenadas de los perfiles renglòn y asi hasta el eje $\underline{m}_K$.

La expresiòn (2.6),pero ahora considerando la variaciòn de los perfiles renglòn respecto a su centroide es

$$\sum_i p_{i.}(\underline{r}_i-\underline{c})'\bar{D}_c^{-1}(\underline{r}_i-\underline{c}) \qquad (2.31)$$

que es igual a

$$\text{traza}[D_r(R-\underline{1}\underline{c}')\bar{D}_c^{-1}(R-\underline{1}\underline{c}')'] \ . \qquad (2.32)$$

Ahora,por (2.17),(2.30) y (2.11) se tiene que

$$\text{traza}[D_u^2]=\text{traza}[F'D_r F]=\text{traza}[D_r FF']$$

$$=\text{traza}[D_r\bar{D}_r^{-1}(P-\underline{r}\underline{c}')\bar{D}_c^{-1}MM'\bar{D}_c^{-1}(P-\underline{r}\underline{c}')'\bar{D}_r^{-1}] \qquad (2.33)$$

$$=\text{traza}[D_r(R-\underline{1}\underline{c}')\bar{D}_c^{-1}MM'\bar{D}_c^{-1}(R-\underline{1}\underline{c}')']$$

pero

$$\bar{D}_c^{-1}MM'\bar{D}_c^{-1}=\bar{D}_c^{-1}$$

ya que

$$M'D_c^{-1}M=I$$

por lo que

$$traza[D_u^2]=traza[[D_r(R-\underline{1}\underline{c}')D_c^{-1}(R-\underline{1}\underline{c}')'] \quad (2.34)$$

En vista de (2.34) la variación total original es recuperada íntegramente en la descomposición en valor singular, en la variación de las coordenadas F.

Analogamente respecto a las coordenadas de los perfiles columna

$$\sum_k u^2_k=traza[G'DcG]=j\ p.j(\underline{c}j-\underline{r})'D_r^{-1}(\underline{c}j-\underline{r}) \quad .(2.35)$$

Por último,el A.C. provee de tres medidas para cuantificar la calidad de la representación gráfica;estas tres medidas son:

i)la calidad global de la representación en l dimensiones,reprentada por

$$(u_1^2+\ldots+u_l^2)/\sum_k^K u_k^2 \qquad l<=K.$$

Usualmente se buscará que con l=1 ó l=2 se tenga la representación gràfica.Se puede interpretar como la proporción de la variación global que representan l dimensiones.

ii)La contribución relativa,que permite saber el ángulo de un perfil con respecto a uno de los ejes principales;de la figura 2.2,en tèrminos de la definición del coseno de un ángulo se tiene que

$$(p_i . f_{ik}^2)/(p_i . d_i^2) = (f_{ik}/d_i)^2 = \cos^2 \theta$$

es decir, si $\cos^2 \theta$ es cercano a 1 el ángulo del perfil i será pequeño con respecto al eje en cuestión.

iii) La contribución absoluta, que cuantifica la aportación de un perfil en la variación total de un eje principal

$$(p_i . f_{ik}^2)/u_k^2 .$$

En la práctica la utilidad de estas medidas es muy importante; respecto a la calidad global, lo que generalmente se desea es que con l=1 ó l=2 se tenga aglutinada la mayor parte de la variación original en la nube de perfiles multivariada ya que la aproximación de las relaciones a través de la gráfica unidimensional o bidimensional será mejor.

En el caso en que la calidad global no sea del todo convincente, es decir que oscile entre el 10% y 20%, las otras dos medidas mencionadas arriba entran en juego; la contribución absoluta identifica a aquellos perfiles que pesaron más en la orientación de cada uno de los ejes principales, mientras que la contribución relativa medirá la correlación de cada perfil con respecto a cada uno de los ejes principales, dando así una idea de que tan bien o que tan mal estuvieron representados. Los puntos

que aparecen cercanos al origen de la gràfica estaràn mal representados,con excepciòn de aquellos que coincidan con su centroide respectivo.

Asi,mientras la calidad global es suficiente para cuando agrupa la mayor parte de la varianza original,las otras dos medidas calificaràn en forma individual a cada perfil en tèrminos de su aporte en la redistribuciòn de la variabilidad en cada uno de los ejes generados por la DVS asi como su correlaciòn con èstos.

Bibliografía.

Greenacre,M.J.(1984)."Theory and Applications of Correspondance Analysis".Academic Press.

Green,E.P. y Carroll,J.D.(1976)."Mathematical tools for Applied Multivariate Analysis".Academic Press.

Tabet,N.(1973).Programa de Análisis de Correspondencias,Parte de tesis Doctoral Universidad de París VI.

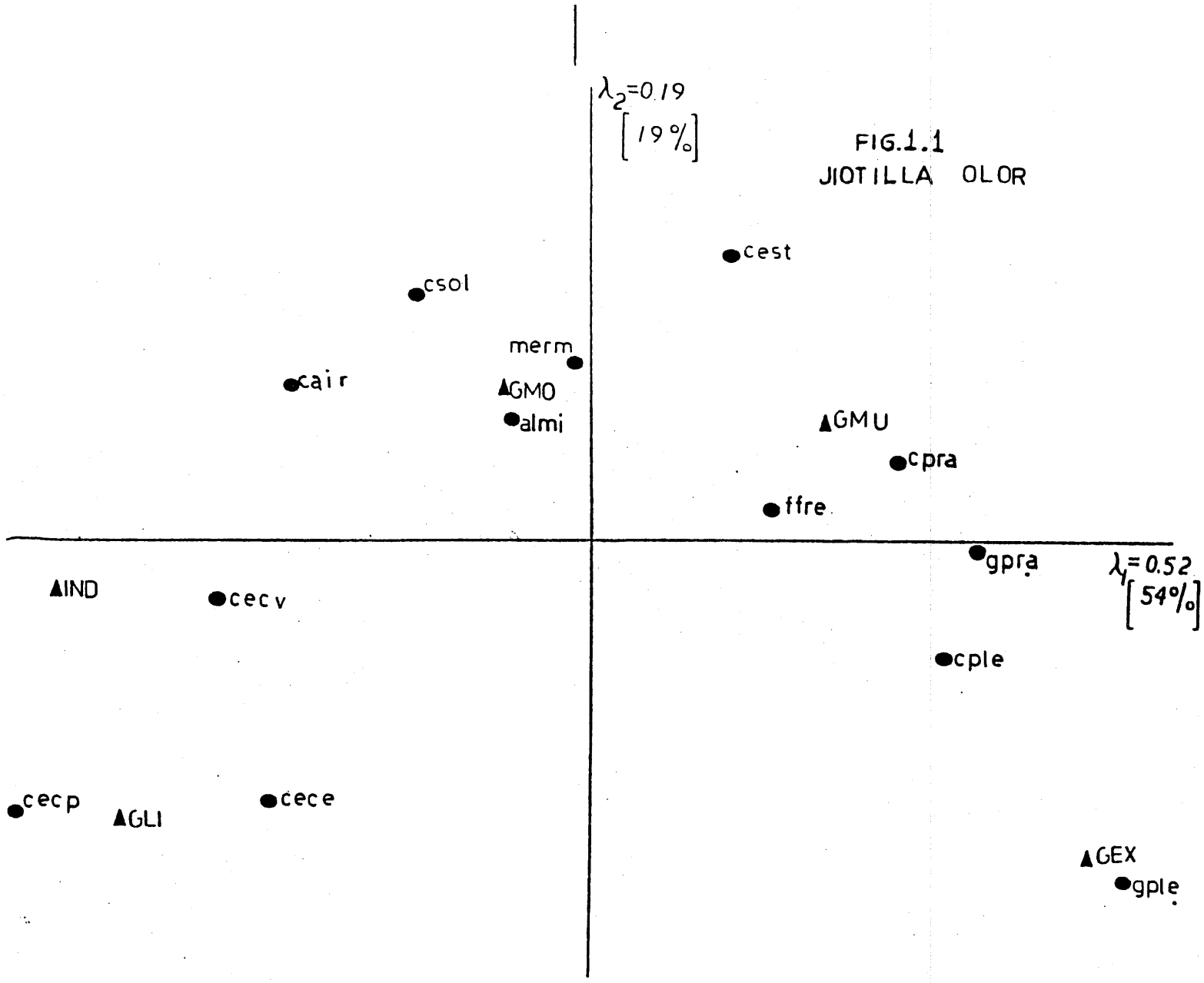


FIG.1.1
JIOTILLA OLOR

FIG 1.2

JIOTILLA SABOR

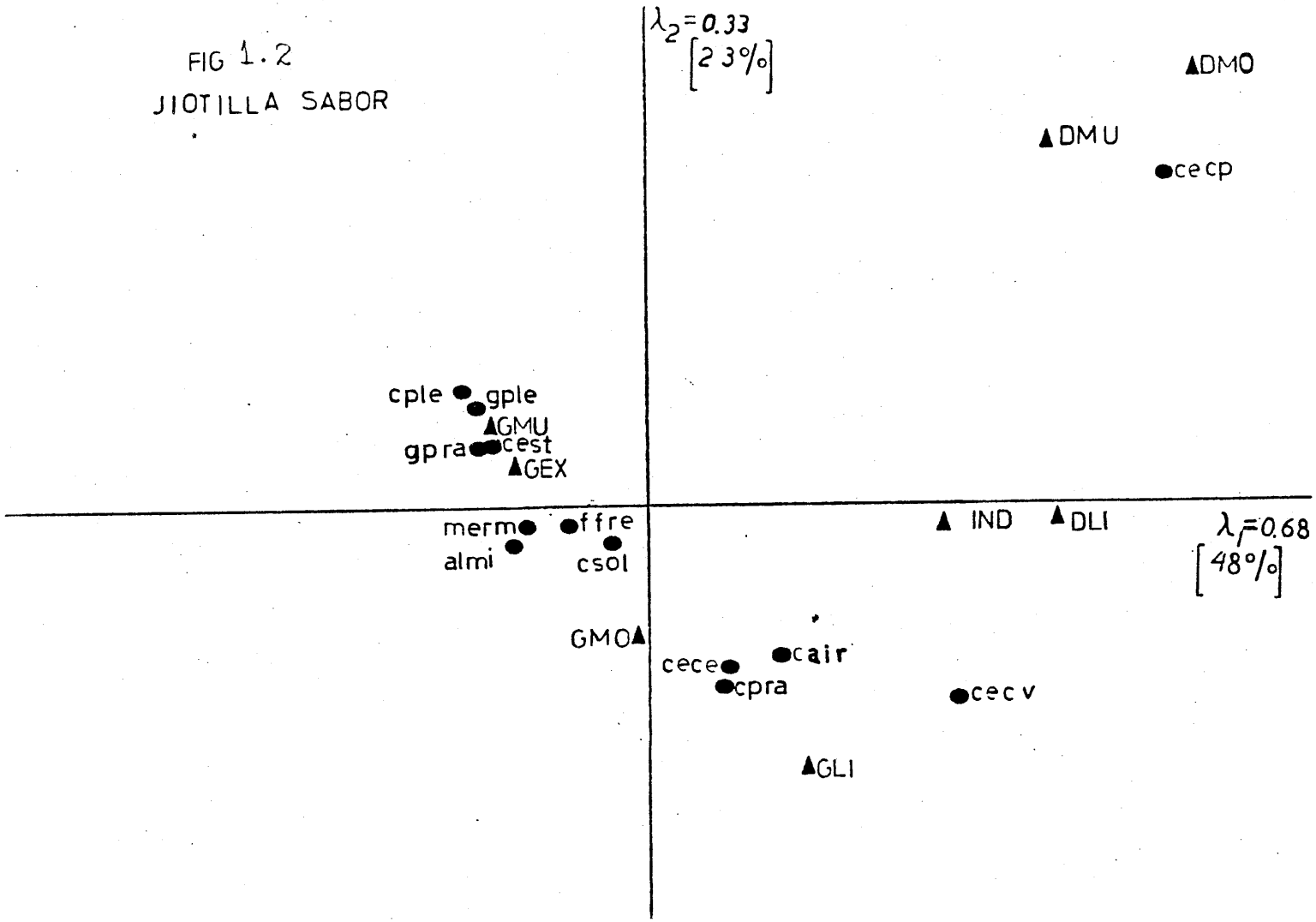


Figura 2.1 Objetivo del Análisis de Correspondencias.

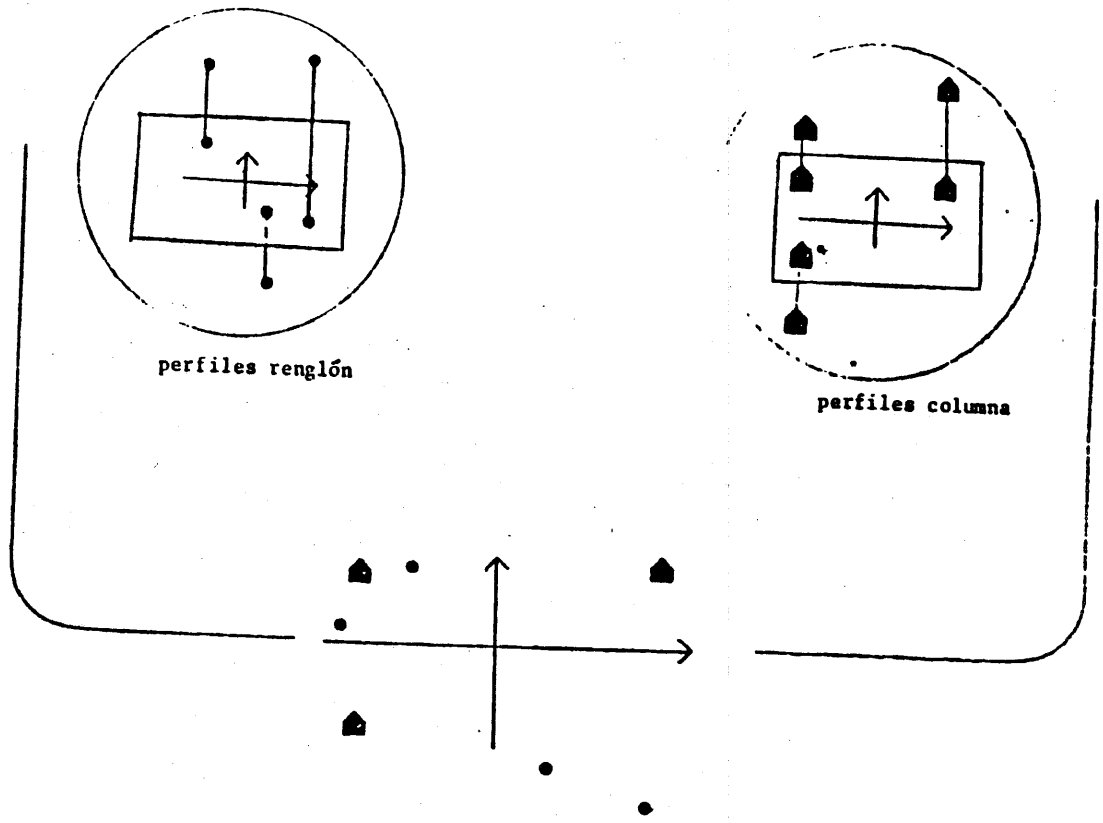
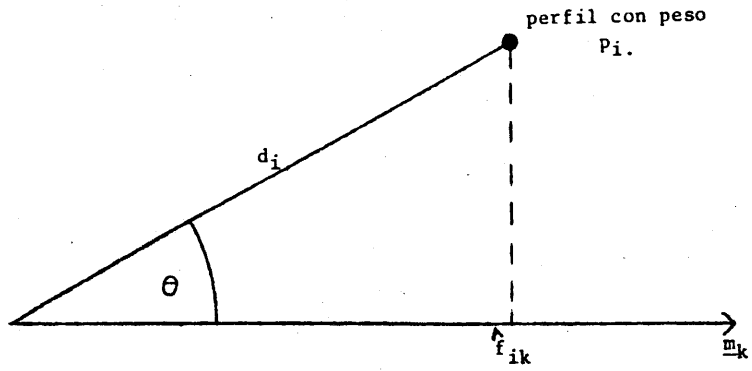


Figura 2.2 Contribuciones relativa(correlación) y absoluta.



Correlación=$(\hat{f}_{ik}/d_i)^2$

Contribución
Absoluta $=(p_i.\hat{f}_{ik})^2/\hat{u}_k^2$

Cuadro 1.1 Tabla de contingencias, OLOR.

	FFRE	CEST	CSOL	CAIP	MERM	ALMI	CPLE	CPRA	GPLE	GPRA	CECP	CECE	CECV
	15.	15.	15.	15.	15.	15.	15.	15.	15.	15.	15.	15.	15.
GEY	4.	0.	0.	0.	0.	0.	6.	3.	9.	4.	0.	1.	0.
GGMJ	3.	9.	0.	0.	0.	0.	5.	3.	6.	10.	0.	1.	0.
GGMO	3.	0.	0.	0.	0.	0.	5.	3.	6.	1.	0.	1.	0.
GLII	3.	0.	0.	0.	0.	0.	5.	3.	6.	0.	0.	1.	0.
IMGI	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
IOLM	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
OOMJ	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
OOFX	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.

Cuadro 1.2.Contribuciones Absolutas y Correlaciones.
Escala de preferencias,OLOR.

	CONTRIBUCIONES ABSOLUTAS		CORRELACIONES	
	eje 1	eje2	eje1	eje2
GEX	0.314	0.381	0.648	0.283
GMU	0.149	0.101	0.604	0.148
GMO	0.024	0.181	0.153	0.417
GLI	0.279	0.276	0.621	0.221
IND	0.223	0.006	0.558	0.005
DLI	0.011	0.055	0.098	0.184

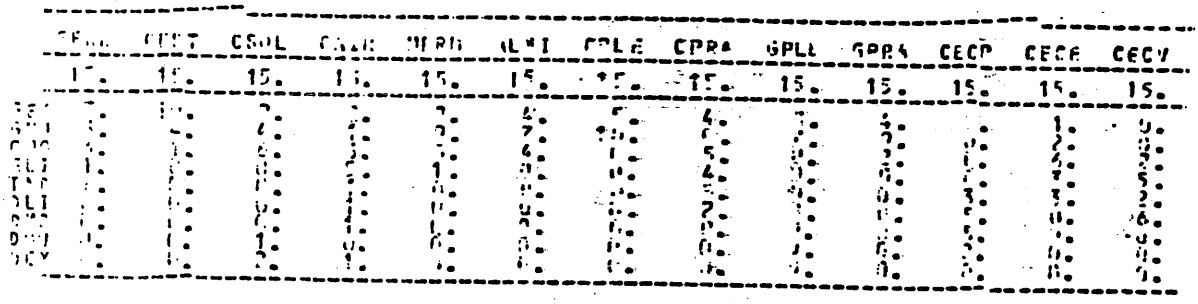
Cuadro 1.3 Contribuciones Absolutas y Correlaciones.
TRATAMIENTOS,OLOR.

	CONTRIBUCIONES ABSOLUTAS		CORRELACIONES	
	eje 1	eje2	eje1	eje2
FFRE	0.023	0.002	0.288	0.010
CEST	0.013	0.167	0.142	0.645
CSOL	0.023	0.118	0.220	0.413
CAIR	0.066	0.046	0.345	0.087
MERM	0	0.061	0.007	0.463
ALMI	0.005	0.026	0.108	0.190
CPLE	0.084	0.031	0.787	0.103
CPRA	0.063	0.010	0.789	0.045
GPLE	0.203	0.241	0.659	0.282
GPRA	0.103	0	0.762	0
CECP	0.240	0.150	0.776	0.175
CECE	0.076	0.141	0.337	0.227
CECV	0.100	0.007	0.848	0.020

Cuadro 1.5 Contribuciones absolutas y Correlaciones.
Escala de preferencias,SABOR.

	CONTRIBUCIONES ABSOLUTAS		CORRELACIONES	
	eje1	eje2	eje1	eje2
GEX	0.110	0.024	0.485	0.050
GMU	0.179	0.089	0.623	0.149
GMO	0	0.147	0.001	0.464
GLI	0.054	0.350	0.214	0.666
IND	0.103	0	0.646	0.001
DLI	0.277	0	0.780	0
DMO	0.219	0.280	0.579	0.356
DMU	0.058	0.099	0.418	0.340
DEX	0	0.011	0.001	0.038

Cuadro 1.4.Tabla de contingencias ,SABOR



Cuadro 1.6 · Contribuciones Absolutas y Correlaciones. Tratamientos, SABOR.

	CONTRIBUCIONES ABSOLUTAS		CORRELACIONES	
	eje1	eje2	eje1	eje2
FFRE	0.012	0.001	0.302	0.010
CEST	0.048	0.020	0.412	0.082
CSOL	0.003	0.03	0.018	0.008
CAIR	0.029	0.084	0.251	0.349
MERM	0.030	0.001	0.454	0.008
ALMI	0.035	0.002	0.702	0.023
CPLE	0.063	0.052	0.564	0.226
CPRA	0.008	0.126	0.084	0.610
GPLE	0.059	0.042	0.585	0.201
GPRA	0.048	0.020	0.674	0.136
CECP	0.489	0.401	0.714	0.281
CECE	0.008	0.103	0.062	0.381
CECV	0.167	0.144	0.527	0.219

AJUSTE DE CURVAS INDIVIDUALES DE CRECIMIENTO HUMANO.

Instituto de Investigaciones Antropologicas, UNAM.

RESUMEN

Un hecho comun en el estudio del crecimiento humano es la estimación de caracteristicas individuales derivadas a partir de observaciones longitudinales. Con tal fin se han propuesto varios modelos, que, generalmente, son no lineales en sus parametros. La estimación de estos ultimos plantea problemas tanto computacionales como referentes a los supuestos involucrados por distintos metodos de estimación. En este trabajo se discuten tales problemas, ilustrandolos con el modelo Preece-Baines 1. Unicamente se presentan los problemas referentes a la metodologia de estimación seguida; los datos, procedentes del estudio de crecimiento realizado en la Universidad de Londres bajo la dirección del Prof. J.M. Tanner, se analizan en mi trabajo de tesis de Maestria en Ciencias, realizada bajo la dirección del Dr. Francisco Aranda Ordaz y que sera presentada proximamente en la UACPyP del CCH/UNAM. Debo la obtención de tales datos a la Mtra. Maria Elena Saenz, del IIA.

1.- CURVAS DE CRECIMIENTO HUMANO.

Durante todo este trabajo el termino crecimiento se referirá al aumento de tamaño manifestado en el cuerpo humano a partir del nacimiento. Este proceso, que en realidad abarca desde el momento de

la concepción hasta que el individuo se convierte en adulto, presenta patrones generales, comunes a la especie humana, en la forma en que ocurren los incrementos de tamaño respecto a la edad cronológica (Faulhaber, 1976:7).

Es posible entonces suponer la existencia de curvas del crecimiento que representen las relaciones que guardan la edad y el crecimiento. En general, estas curvas tendrán una forma similar para las poblaciones humanas así como para los individuos que las constituyen. Dentro de esta unidad en la forma en que ocurre el fenomeno, existen variaciones (entre poblaciones y entre individuos) condicionadas esencialmente por factores hereditarios (influencias internas) y ambientales (influencias externas). Puede afirmarse que es sumamente difícil separar los efectos de estos factores sobre el crecimiento. En vista de lo anterior, resulta razonable preguntarse por modelos matemáticos que representen la forma general del proceso de crecimiento, dando cuenta de sus posibles variaciones a partir de medir variables relativamente fáciles de observar.

La finalidad de construir y ajustar modelos de crecimiento es doble: por una parte, es posible condensar adecuadamente la información contenida en volúmenes relativamente grandes de datos; por otra, se tienen mejores condiciones para estudiar las relaciones que durante el desarrollo del proceso tiene la edad con respecto a ciertos parámetros (no siempre estimables directamente) que reflejan hechos sobresalientes acerca del fenómeno. Muchas veces, un estudio de crecimiento está centrado en el análisis de la variación individual en la forma en que ocurren los incrementos de crecimiento al interior de una población. Para conseguir esto en una forma exacta, es necesario

observar el crecimiento de los individuos en diferentes momentos; es decir, el diseño de la investigación deberá ser, en algun sentido, longitudinal. Esto permite, ademas, la construcción de estándares para velocidad de crecimiento que pueden utilizarse ya como método para detectar niños que si bien "no están abiertamente enfermos ... podrían obtener atencion especial, ya sea médica, educacional o social" (Tanner, 1978), i. en la práctica pediátrica, como método de estudio para verificar si niños que han estado enfermos reaccionan a la aplicación de algun tratamiento (Berkey *et al.*, 1983). Cabe insistir en que la necesidad de tener estudios longitudinales para lograr adecuadamente lo arriba expuesto se basa en que estos son la única manera de observar las diferencias que existen entre los patrones de crecimiento al interior de la población.

Por otra parte, debido a la no linealidad en los parámetros que presenta la mayoría de los modelos propuestos, hay una diferencia entre la curva que se obtendría promediando las estimaciones individuales y la que debe ser considerada como curva promedio para toda la poblacion, por lo que la comparación de crecimientos individuales con la tendencia general de la población no es adecuada utilizando la curva primeramente mencionada; además, si bien este problema ha sido estudiado para modelos que son lineales en los parámetros (Fearn 1975), es sólo hasta muy recientemente (Berkey y Laird, 1986; Goldstein, 1986) que se ha propuesto una metodología similar para tratar el caso no lineal. En este trabajo únicamente se consideraran curvas de crecimiento individual aplicables, en principio, a datos longitudinales.

El uso de modelos matemáticos en el estudio del crecimiento se ha

dado bajo dos enfoques que, en cierta forma, resultan antagónicos (Bock y Thissen, 1980; Preece, 1978). Por una parte, el enfoque no estructural intenta aproximarse al análisis de datos de crecimiento a partir de suavizar global o localmente las variaciones individuales observadas en los incrementos. Esto se hace, generalmente, con polinomios de Lagrange (Zerbe, 1979) o bien suavizando globalmente mediante splines cúbicos (Largo *et al.*, 1978). Es importante hacer notar que los parámetros de tales funciones difícilmente tienen una interpretación en términos de aspectos relevantes para el crecimiento; además, en ambos casos no hay consideración estocástica alguna para los errores. Por el contrario, en el enfoque estructural se tiene una función, que depende de la edad y de un conjunto de parámetros, con la que se intenta representar la tendencia general del crecimiento junto con una serie de supuestos aleatorios asociados con las variaciones debidas, por ejemplo, a toda clase de errores de medición, a la inexactitud habida en la representacion matemática de fenómeno, o a cambios en la salud o en las condiciones ambientales del niño. Las funciones propuestas sintetizan relaciones entre ciertos aspectos generales del fenomeno que son conocidas de cierto. Por ejemplo, se sabe de la existencia de una edad en la que la velocidad de crecimiento es máxima durante la adolescencia, de una estatura final fija, de una aceleración negativa del crecimiento durante los primeros años de vida, etc.

Casi siempre, las relaciones entre estos hechos y la edad pueden expresarse con sistemas de ecuaciones diferenciales que al resolverse generan el modelo de crecimiento, el cual resulta, en la mayoría de los casos, no lineal en los parámetros. Se supone que tal modelo será adecuado para algun rango de edades relativamente amplio. Desde

luego, dependiendo de qué tan fielmente refleje el sistema de ecuaciones diferenciales a las características conocidas del fenómeno, el modelo producirá una mejor forma para el patrón de crecimiento; por ejemplo, el modelo logístico supone que el cambio en la aceleración que ocurre en la adolescencia es simétrico, lo que es un error. Por otra parte, al abarcar rangos más o menos grandes, es difícil considerar todos los detalles locales que puede presentar el patrón de crecimiento: tal es el caso del modelo Preece-Baines 1 (Preece y Baines, 1979.), que si bien cubre edades desde el nacimiento hasta la edad adulta, no detecta una pequeña aceleracion del crecimiento (llamada *mid-growth spurt*) que se presenta alrededor de los 7 años. Finalmente, tales modelos no involucran, salvo en la parte aleatoria, la variación atribuible a otras características individuales (ambientales, sociales, estacionales, etc.) que varían respecto al tiempo y, que sin embargo, pueden ser medidas con cierta precisión.

A pesar de estas limitaciones, el enfoque estructural resulta generalmente preferible al no estructural. Algunos ejemplos de modelos estructurales son: el de Jenss-Bayley, para tratar con edades entre 0 y 6 años (Berkey, 1982); el de Gompertz y el logístico (Marubini, 1979), utilizados para estudiar el crecimiento durante la adolescencia; los modelos doble y triple logístico (Bock y Thissen, 1980) y la familia de modelos de Preece y Baines (*op.cit.*), los cuales abarcan desde el nacimiento hasta la edad adulta.

2.- MODELO PREECE - BAINES 1.

Puesto que los modelos propuestos para estudiar el crecimiento desde

el nacimiento hasta la edad adulta no resultaron satisfactorios (por ejemplo, el triple logístico tiene 10 parámetros por estimar, además de que se necesita conocer la estatura adulta para estimarlo), Preece y Baines propusieron una familia de modelos para el ciclo completo de crecimiento que ha probado ser muy superior a otras alternativas. Tal superioridad no es sólo en términos de parsimonia, ya que los parámetros involucrados en el modelo son interpretables biológicamente y a partir de ellos pueden estimarse otras características importantes del fenómeno. La construcción de la familia de modelos Preece - Baines está expuesta con detalle en Preece y Baines (1978).

De los modelos de esta familia, el llamado PB 1 ha sido el más utilizado, además de tener sólo 5 parametros (uno menos que el resto de los modelos PB) y haber producido errores de estimación a lo más del mismo orden que los otros modelos (Preece y Baines, op. cit.).

Este modelo se expresa como:

$$h(t) = h_1 - \frac{2(h_1 - h_0)}{\exp[s_0(t - \theta)] + \exp[s_1(t - \theta)]}, \qquad (1)$$

siendo h_1 la estatura adulta, h_0 un parámetro relacionado con la estatura en el momento de máxima velocidad de crecimiento durante la adolescencia, θ es tal que $h(\theta) = h_0$, s_0 y s_1 son tasas de crecimiento constantes referentes al inicio y al fin de la aceleración del crecimiento habida en la adolescencia.

El énfasis dado por los parámetros del modelo a la región de la adolescencia no es casual: obedece a que ésta es la parte del ciclo de crecimiento que resulta más difícil de modelar.

Las gráficas 1, 2 y 3 muestran las funciones de distancia, velocidad y aceleración para ambos sexos construidas con la ecuacion (1) y sus dos primeras derivadas respecto al tiempo.

2.- METODOS DE ESTIMACION.

Dado un modelo estructural para la talla, junto con un conjunto de supuestos estocásticos para los errores y n observaciones, la forma usual de plantearlo es como un modelo de regresión no lineal (Gallant, 1975); es decir,

$$y_t = f(x_t; \theta) + e_t \qquad t = 1, 2, \ldots, n \qquad (2)$$

donde y_t es la estatura en la edad x_t, f es la función de crecimiento propuesta, θ es un vector p dimensional de parámetros y las e_t's representan errores aleatorios.

Si f fuera lineal en θ, las ecuaciones (1) podrian escribirse como

$$y_t = \sum_{k=1}^{p} x_t \theta_k + e_t ,$$

o, en notación vectorial,

$$\mathbf{y} = \mathbf{X}\theta + \mathbf{e},$$

siendo $\mathbf{X}$ una matriz de $n \times p$, llamada matriz diseño; entonces, los estimadores de Mínimos Cuadrados (en adelante MC) para θ cumplen con la condición de minimizar la forma cuadrática

$$\Phi(\theta) = (y - X\theta)' (y - X\theta) = \sum_{t=1}^{n} e_t^2 , \qquad (3)$$

lo que se consigue, usualmente, resolviendo el sistema de ecuaciones lineales

$$\theta = (X'X)\, X'y.$$

Es importante hacer notar que los estimadores MC no consideran supuesto alguno respecto a la distribución de los errores.

En el caso no lineal, los estimadores MC se obtienen minimizando la expresión analoga a (3):

$$\Phi(\theta) = \sum [\, y_t - f(x_t; \theta)\,]^2. \qquad (4)$$

Generalmente este problema no admite una solución analítica, por lo que se resuelve por métodos numéricos. A pesar de que ésto puede traer problemas que el caso lineal no presenta (selección de valores iniciales, condiciones de tolerancia y criterios de convergencia adecuados), en caso de contar con una buena rutina para minimizar funciones no lineales , una computadora precisa y algun conocimiento, así sea muy somero, acerca del comportamiento de θ, la solución al problema anterior es relativamente simple. En este trabajo, se utilizaron las rutinas NL2SOL (Calderon, *et al.*, 1983) y CONMIN (Calderon, 1985), asi como una rutina con el método Newton Raphson para resolver los problemas de minimización de sistemas de ecuaciones no lineales no lineales con varios parámetros. Los programas necesarios se codificaron en FORTRAN, con variables de presición

simple en la Burroughs B7900 de la Dirección General de Cómputo Académico de la UNAM.

Puede afirmarse que los problemas más serios de la estimación en modelos no lineales están más bien en lo intratable que casi siempre resulta el desarrollo analítico de sus propiedades estadísticas.

Como es bien sabido, en los modelos lineales, los estimadores MC son los mejores estimadores lineales insesgados y se distribuyen asintóticamente normales; por otra parte, en el caso de que los errores sean normalmente distribuidos, los estimadores MC tienen esta distribución y coinciden con los estimadores de Máxima Verosimilitud (en adelante MV). Como señala Ratkowsky (1983:41.4), si los errores son no normalmente, los estimadores MC no lineales tienen las propiedades del caso lineal asintóticamente. La diferencia fundamental estriba en que el espacio de soluciones de los estimadores MC para el caso no lineal es no lineal. Mientras menos no lineal sea tal espacio, el comportamiento de los estimadores se aproximará más rápidamente a las propiedades anteriores. Si el modelo es no lineal en θ, las posibles reparametrizaciones podrán cambiar drásticamente la forma del espacio de soluciones; además, hay una no linealidad intrínseca al modelo reflejada en la curvatura del espacio. De lo anterior se desprende la importancia de establecer el grado de no linealidad en la estimación del modelo con el que se está trabajando. Aunque desde 1960 Beale propuso ciertas medidas de no linealidad, modificadas por Box en 1971 y por Bates y Watts en 1980, este aspecto de la estimación ha sido poco trabajado; de hecho, no ha sido utilizado en los estudios de curvas de crecimiento humano que fueron consultados. En este trabajo se utilizó como tal medida la estimación

del sesgo del estimador propuesto por Box (1971: § 2). La forma de calcularlo es:

$$S(\theta) = -\tfrac{1}{2}\sigma^2 \left[\sum_{t=1}^{n} F_t F_t'\right]^{-1} \sum_{t=1}^{n} F_t \; tr\left\{\left[\sum_{t=1}^{n} F_t F_t'\right]^{-1} H_t\right\} \quad (5)$$

con $F_t = \nabla f(x_t; \theta)$ y H_t el Hessiano de $f(x_t;\theta)$; en la práctica, σ^2 y θ se sustituyen por sus estimadores. Ratkowsky (*op. cit*: 21) menciona que es adecuado medir la no linealidad de los estimadores con el porcentaje de sesgo para cada componente del parámetro, añadiendo que si éste es menor del 1% indicará que el comportamiento de los estimadores es bastante parecido al que se tiene en el caso lineal. Otra forma de medir la no linealidad es lo cercano a cero que esté la suma de residuales (por ser 0 en el caso lineal) y la singularidad de la matriz de varianza covarianza de los residuales: si ésto ocurre (o casi), se tendrá un bajo grado de no linealidad en el espacio de soluciones, puesto que con el caso lineal, ésto pasa, debido a que el rango de X es, a lo más, p ($p < n$) y tal matriz es de $n \times n$. Una consideración importante es la hecha recientemente por Cook y Tsai (1985) en el sentido de que si no hay una linealidad aceptable para el espacio de soluciones, será necesario utilizar una clase de residuales generalizados para el análisis del ajuste.

Además de estas consideraciones sobre la no linealidad de la estmación MC, hay un problema concerniente a las curvas individuales de crecimiento: como se mencionó, los estimadores MC no lineales tendran asintóticamente un comportamiento similar a los lineales si el sesgo es pequeño y si los errores son *iid* normales. Asi, para datos longitudinales de crecimiento, además de que es sumamente

difícil hacer que n sea muy grande (lo que, finalmente, podría no ser crucial), este último supuesto difícilmente se cumpla ya que, dependiedo de qué tan separadas están las observaciones, existirá dependencia en la variación residual.

Si se tienen n observaciones equiespaciadas para m individuos y se tiene que las mediciones fueron realizadas en las mismas edades para todos ellos, entonces la forma más natural de incorporar en la estimación tal comportamiento de los residuales es utilizar el metodo de MC generalizados, utilizando como matriz de ponderaciones a la inversa de la matriz de varianza covarianza de los residuales. Como muestra Bard (1974:64), bajo el supuesto de normalidad para los residuales, este método es equivalente al de MV, por lo que así se le designara de aqui en adelante.

Suponiendo que $e \sim N_n (0, \Sigma_e)$, la funcion de log verosimilitud puede escribirse como:

$$\lambda(\theta) = C - \frac{1}{2} \sum_{j}^{n} \sum_{k}^{n} \sigma^{jk} [y_j - f(x_j; \theta)] [y_k - f(x_k; \theta)] \qquad (6)$$

siendo C una constante independiente de θ y σ^{jk} el elemento j-k de Σ_e^{-1} (Bock y Thissen, op. cit.:246). Una vez planteada la ecuación anterior, los estimadores MV se obtienen maximizándola, lo que es equivalente a minimizar $-\lambda$, por lo que esa parte no tiene problema. Lo malo es que en (6) se supone que Σ_e es conocida, y ésto no es asi. Bock y Thissen (1980:271) afirman que, puesto que las σ^{jk} unicamente juegan el papel de ponderaciones, no es necesario que sean estimadas con gran precisión. Siguiendo a Fearn (1975), quien trata un problema de estimación en un modelo lineal para datos longitudinales, estos autores

plantean la siguiente estrategia para obtener estimadores MV:

1) Obtener estimadores MC para las curvas de m individuos. Estos estimadores pueden denotarse por $\theta^{\dagger}_{i}$ (i = 1, ... m).

2) Con estos valores, calcular los residuales

$$e^{\dagger}_{it} = y_{it} - f(x_t; \theta^{\dagger}_{i}), \quad (t = 1, ..., n; i = 1, ..., m),$$

para cada individuo.

3) Estimar, a partir de $e^{\dagger}_{it}$ la matriz Σ_e, con:

$$\sigma^{\dagger}_{jk} = 1/m \cdot \sum_{i=1}^{m} e^{\dagger}_{ij}\, e^{\dagger}_{ik}$$

4) Obtener Σ_e^{-1} y calcular los estimadores MV a partir de (5).

5) Volver al paso 1), ahora con los residuales obtenidos con los estimadores MV del paso 4)

Bock y Thissen afirman que basta realizar el procedimiento arriba descrito en dos pasos para tener estimaciones "suficientemente exactas". de σ^{jk}

Como se mencionó, si el espacio de soluciones es casi lineal, la matriz Σ_e será casi singular. El hecho de que Bock y Thissen no reporten problemas para obtener la inversa de esta matriz únicamente apunta la alta no linealidad de las estimaciones para el modelo triple

logístico.

Otra posibilidad de especificar mejor de la correlación existente entre los residuales consiste en introducir supuestos acerca de la estructura de ésta. Lo más simple es suponer que los residuales para cada individuo como generados por un proceso estocastico estacionario normal, es decir, con media y varianza independientes del tiempo, de tal forma que la correlacion entre los residuales para las edades x_j y x_k unicamente dependa de la distancia habida entre x_j y x_k. De esta forma se evita el problema del calculo de Σ_e^{-1} cuando la estimación es parecida al caso lineal. Glasbey (1979) trata el problema con el modelo logístico generalizado. Por otra parte, Gallant y Goebel (1976) proponen otra estrategia, aplicable cuando el número de observaciones es tan grande que podría plantear serios problemas para almacenar e invertir a Σ_e; fundamentalmente, esta enfoque se basa en especificar el orden del proceso estacionario que se supone para los residuales.

En este trabajo, se consideró a la matriz de varianza covarianza de los residuales como una matriz Toeplitz Γ_n de n×n, (Hannan, 1993:27), estimando a cada componente de esta matriz como el promedio sobre los m individuos de:

$$\gamma_{ik} = 1/(n-p) \sum_{j=1}^{n-k} e_{ij}\, e_{ij+k}$$

De esta forma, Γ_n^{-1} fue utilizada como Σ_e^{-1} en la ecuacion (6).

Los estimadores MV se obtuvieron con el metodo de Newton Raphson (Kalbfleisch, 1979; Silvey, 1975), utilizando como valores iniciales en el primer paso a los estimadores MC y posteriormente, a los estimadores MV calculados en el paso anterior. Este

procedimiento se efectuó 10 veces con el fin de examinar cuándo era posible considerar que la matriz de varianza covarianza de los residuales permanecía estable. Analizando tanto las gráficas de iteraciones promedio como del condicional de las matrices circulares mencionadas arriba contra el número de pasos se determinó que 4 pasos eran más adecuados para los estimadores MV. La matriz de varianza covarianza fue actualizada en cada paso.

Otra justificación para el uso de estimadores MV es el principio de invarianza, que afirma que el estimador MV de una función del parámetro es la función evaluada en el estimador MV. Esto era importante en este trabajo ya que algunos parámetros biológicos son funciones de los parámetros. Cabe mencionar que para el cálculo de dichos parámetros es frecuente tener que encontrar las raíces de ciertas funciones no lineales; por ejemplo, para encontrar la edad en la que la velocidad de crecimiento es máxima durante la adolescencia, es necesario saber los puntos del tiempo en los que se anula la segunda derivada (respecto a la edad) de la función de crecimiento. Algunas rutinas aplicables para esta clase de problemas aparecen en Calderón (1995).

Finalmente, se trabajó con el método Bayesiano Empírico (Maritz, 1970) (en adelante, BE), estimándose la distribución *a priori* con los estimadores MV obtenidos para cada individuo. Utilizando los índices $\sqrt{b_1}$ y b_2 de D'Agostino y Pearson (1973) fue posible establecer las normalidades marginales para los cinco componentes de tal distribución. Si $\theta \sim N_p(\mu_\theta, \Sigma_\theta)$, la función de log verosimilitud puede escribirse como

$$\phi(\theta) = D - \frac{1}{2}\sum_{j}^{n}\sum_{k}^{n}\sigma^{jk}\,a_j\,a_k - \frac{1}{2}\sum_{g}^{S}\sum_{h}^{S}\sigma_{\theta}^{gh}\,(\theta_g - \mu_g)\,(\theta_h - \mu_h), \quad (7)$$

siendo D una constante que no depende de θ y σ_{θ}^{gh} elemento de Σ_{θ}^{-1}. Notese que la ecuación de máxima verosimilitud necesaria para el procedimiento Newton-Raphson es la ecuación anterior sin el sumando correspondiente a la distribucion *a priori*. La utilidad de este método está en que permite realizar estimaciones razonables aun con n bastante pequeña; incluso n puede ser 1. Por esta razón, el método puede ser muy útil en sistemas de información clínica en los que se podría estimar los parámetros del crecimiento de un individuo, una vez asignadas matrices de varianza covarianza para los parámetros y para los residuales, desde la primera vez que se le observa, actualizando tales estimaciones con las siguientes mediciones.

Otro enfoque seguido fue el de analizar directamente la función de log verosimilitud relativa (Edwards, 1984), intentando establecer regiones de plausibilidad. Sin embargo, el hecho de que el parámetro fuera de dimensión 5 dificultó este análisis. La idea consistió en fijar cada uno de los componentes del parámetro en el valor de su estimador MV y posteriormente variar sobre algun rango razonable (determinado con base en la inversa de la matriz de información esperada calculada para cada individuo) el resto de los componentes, conjuntamente. Los resultados obtenidos no fueron de utilidad para determinar las regiones de plausibilidad; en general, se observó que la información estaba sumamente concentrada en vecindades del estimador MV que resultaron ser demasiado pequeñas para ser de alguna utilidad práctica en cuanto a establecer tales regiones. En cierta forma, ésto habla bien de los

estimadores MV obtenidos, pues se puede afirmar que la información que la muestra contiene acerca del parámetro (medida con la función de verosimilitud) está muy concentrada alrededor del estimador MV (llamado, en el contexto que maneja Edwards, evaluador).

Cabe mencionar que se calculó, para cada individuo, la matriz de informacion esperada con el fin de obtener una estimación de la variabilidad del estimador.. Esto permitió construir intervalos de confianza aproximados, utilizando una distribución t de Student, para cada componente del parámetro. Al respecto, véase el articulo de Gallant y Goebel (1976).

Los resultados obtenidos con los tres métodos fueron prácticamente iguales, con excepción de dos individuos que tuvieron valores muy diferentes al resto de la muestra en el estimador de s_1 para el metodo BE. Era de esperar que este parámetro (o bien s_0) fuera el que peor comportamiento podía tener, puesto que se puede intercambiar con s_0 sin alterar el modelo. Respecto a la estimación BE realizada con una sola medición, se intentó considerando como la edad en la que la única observación era realizada a los 5, 9 y 14 años. Para la primera, no fue posible ajustar el modelo en la mayoría de los individuos, obteniéndose ajustes bastante malos. En las otras dos edades se tuvieron ajustes satisfactorios para casi todos los individuos.

Puede concluirse que, en general, si el sesgo de Box para el estimador resulta relativamente pequeño, y si no hay una fuerte evidencia de autocorrelación para los residuales (medida en este trabajo con la prueba de rachas y con la estadística de Durbin-Watson), los estimadores MC sin restricciones funcionarán casi tan bien (o mejor) que los estimadores MV o BE, siendo, además, mucho más

fáciles de calcular. En otros casos, será necesario cambiar la estrategia de estimación. Sin embargo, no se encontró mucho énfasis en estos problemas (en particular a los relacionados con las medidas de no linealidad para el espacio de soluciones) en la literatura consultada para el trabajo de tesis.

REFERENCIAS

Bard, Y.; (1974). Nonlinear parameter estimation, Academic Press, Nueva York.

Bates, D.M. y D.G. Watts; (1980). "Relative curvature measure of nonlinearity (with discussion)", *JRSS, ser. B*, 42, 1-25.

Beale, E.M.L.; (1960). "Confidence regions in nonlinear estimation", *JRSS. ser. B*, 22, 41-76.

Berkey, C.S.; (1982). "Comparison of two longitudinal growth models for preschool children", *Biometrics*, 38, 221-234.

Berkey, C.S., R.B. Reed e I. Valadian; (1983). "Longitudinal growth standards for preschool children", *Ann. Hum. Biol.*, 10, 57-67.

Berkey, C.S. y N.M. Laird; (1986). "Non linear growth curve analysis: estimation of the population parameters", *Ann. Hum. Biol.* 13, 111-128.

Bock, R.D. y D. Thissen; (1980). "Statistical problems of fitting individual growth curves", en Johnston, F.E. y A.F. Roche (eds.), Human physical growth and maturation: methodologies and factors, 265-290, Plenum, Nueva York.

Box, M.J.; (1971). "Bias in non linear estimation", *JRSS, ser. B*, 33, 171-201.

Calderón, A.; (1985). Guía par el uso de la biblioteca básica de programas de análisis numérico. Parte I, *Comunicaciones Técnicas, serie azul*, 50, IIMAS, UNAM, México.

Calderón, A, S. Gómez, J. Alonso, J. Capistrán, P. Guerrero, J.L. Morales y M. Sánchez; (1983). NL2SOL, una subrutina altamente robusta para resolver problemas de mínimos cuadrados no lineales, *Comunicaciones Técnicas, serie azul*, 69, IIMAS, UNAM, México.

Cook, R.D. y C.L. Tsai; (1985). "Residuals in nonlinear regression", *Biometrika*, 72, 23-29.

D'Agostino, R. y E.S. Pearson; (1973). "Tests for departure from normality. Empirical results for the distributions of b_2 and $\sqrt{b_1}$", *Biometrika*, 60, 613-622.

Edwards, A.W.F.; (1984). Likelihood, Cambridge University Press.

Faulhaber, J.; (1976). Estudio longitudinal del crecimiento, Instituto Nacional de Antropología e Historia, México.

Fearn, T.; (1975). "A bayesian approach to growth curves", *Biometrika*, 64, 89-100.

Gallant, A.R.; (1975). "Non linear regression", *Am. Statistician* 29, 73-81.

Gallant, A.R. y J.J. Goebel; (1976). "Non linear regression with Autocorrelated errors", *JASA*, 71, 961-967.

Glasbey, C.A.; (1979). "Correlated residuals in non-linear regression applied to growth data", *App. Statist.* 28, 251-259.

Goldstein, H.; (1986). "Efficient statistical modelling of longitudinal data", *Ann. Hum. Biol.*, 13, 129-141.

Hannan, E.J.; (1983). Time series analysis, Chapman and Hall, Londres.

Kalbfleisch, J.G.; (1979). Probability and Statistical Inference, Springer-Verlag, Berlin.

Largo, R.H., T. Gasser, A. Prader, W. Stutzle y P.J. Huber; (1978). "Analysis of the adolescent growth spurt using smoothing spline functions", *Ann. Hum. Biol.*, 5, 421-436.

Maritz, J.S.; (1970). Empirical Bayes methods, Methuen, Londres.

Marubini, E.; (1978). "The fitting of longitudinal growth data of man", en Gedda, L. y P. Parisi (eds.), Auxology: Human growth in health and disorder, 121-131, Academic Press, Londres.

Preece, M.A.; (1978). "Analysis of the human growth curve",

Postgraduate Medical Journal, suppl. 1, 54, 77-86.

Preece, M.A. y M.J. Baines: (1978). "A new family of mathematical models describing the human growth curve", *Ann. Hum. Biol.* 5, 1-24.

Ratkowsky, D.A.: (1983). Nonlinear regression modelling. Dekker, Nueva York.

Silvey, S.D.; (1975). Statistical Inference, Chapman and Hall, Londres.

Tanner, J.M.; (1979). "Human growth standards: construction and use", en Auxology: Human growth in health and disorder, 109-121.

Zerbe, G.O.; (1979). "A new nonparametric technique for constructing the percentiles and normal ranges for growth curves determined from longitudinal data", Growth, 43, 283-278.

Ciudad Universitaria, D.F.,
octubre de 1986.

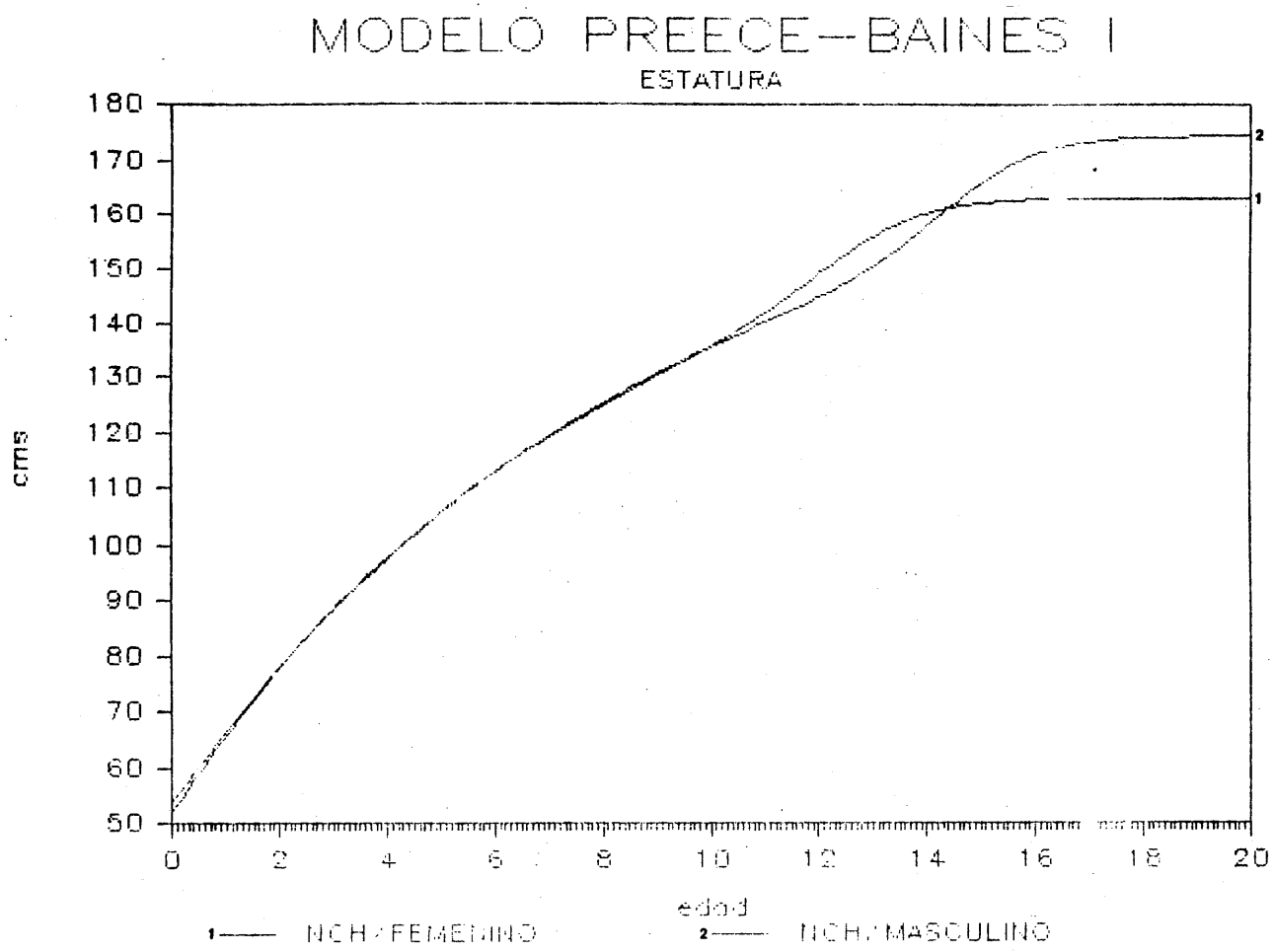

MODELO PREECE-BAINES I
ESTATURA
cms
180
170
160
150
140
130
120
110
100
90
80
70
60
50
0
2
4
6
8
10
12
14
16
18
20
edad
1 —— NCH. FEMENINO
2 —— NCH. MASCULINO

FIG. 2

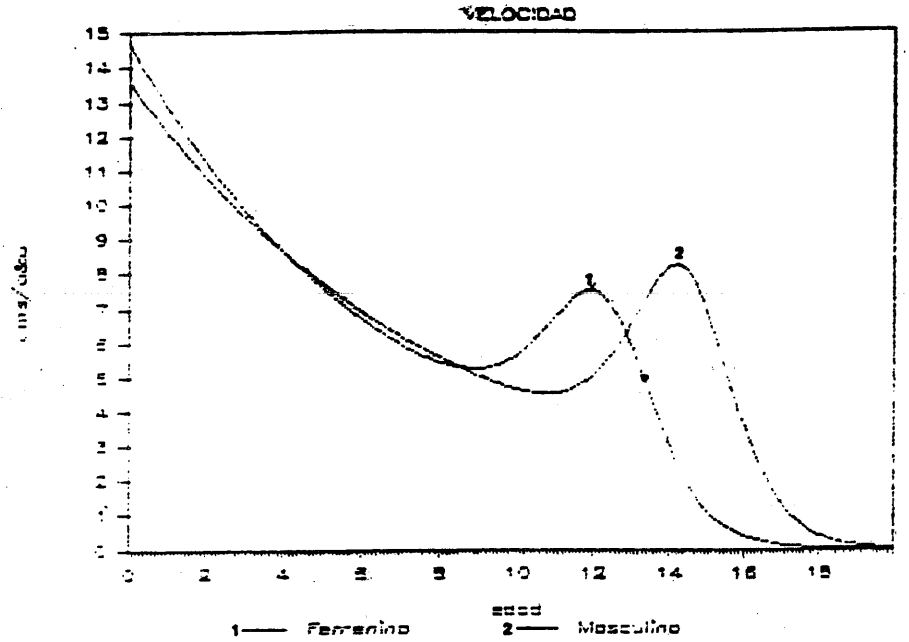
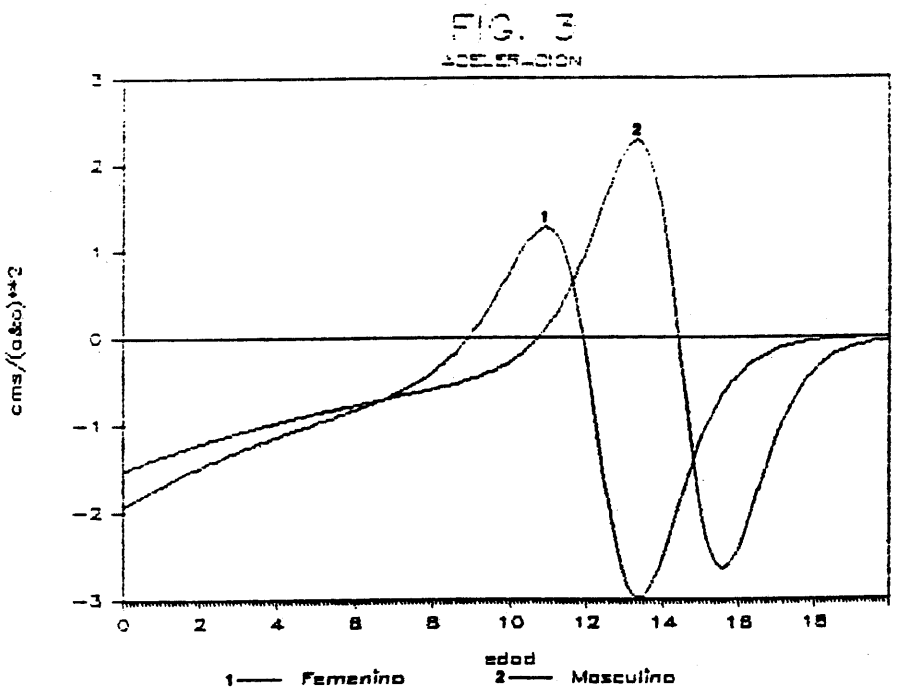

VELOCIDAD
edad
1—— Femenino
2—— Masculino

FIG. 3


ACELERACION
cms/(a&o)**2
edad
1—— Femenino
2—— Masculino

PRESENTACION DE UN PAQUETE ESTADISTICO RELACIONADO CON MODELOS NO LINEALES

Antecedentes

El Instituto Nacional de Nutrición realizó una investigación en el Edo. de Puebla para determinar el grado de desarrollo físico de los niños, desde recien nacidos hasta los seis años, cuando son sometidos a un regimen de alimentación. Consideraron como variables de respuesta el peso y estatura de los niños,como variables de entrada la edad.

Con los datos obtenidos se proporcionan los siguientes modelos no lineales:

$$Y=\beta_1(1-\exp(\beta_2-\beta_3 X)),$$

$$Y=\beta_1-\exp(-\beta_2)\beta_3^{x}.$$

En el laboratorio de cultivo de tejidos vegetales del Instituto de Biología, se analizó la respuesta in vitro de Capsicum chinense (chile habanero) y Capsicum annum L.serrano. Se hicieron ciertos cortes para estudiar el crecimeinto de callo del hipocótilo en un medio cultivo.Se consideró como variable de respuesta el peso en seco obtenido en el tiempo t. Con los datos obtenidos se propusieron los modelos siguientes:

Richards $Y=\alpha/(1+\exp(\beta-\tau X))^{1/\delta}$.

Morgan-Mercer-Flodin $Y=(\beta\tau-\alpha x^{\delta})/(\tau+X^{\delta})$.

Ante este tipo de situaciones, generalmente se requiere ajustar un modelo a un conjunto de observaciones generadas de la investigación, ó considerar un modelo no anidado que mejor explique el comportamiento de los datos, esto da lugar a tener un o más modelos, entonces se plantea una prueba de hipótesis sobre los modelos que se propongan. Con el objeto de resolver estos problemas se propone el paquete estadistico REGROB, que a continuación se describe.

• PAQUETE REGROB

1. INTRODUCCION.

La idea esencial de esta exposición es la de presentar un paquete estadístico que hemos denominado REGROB el cuál puede ser utilizado en las siguientes situaciones:

i) Ajustar un modelo no lineal a un conjunto de datos por mínimos cuadrados o el método robusto.

ii) Realizar pruebas de hipótesis sobre los parámetros,usando la prueba de Wald.

iii) Hacer pruebas de hipótesis sobre modelos no anidados.

La cuestión a desarrollar en el trabajo es hacer un resumen formulando el problema en cada una de las situaciones mencionadas anteriormente. Detallar lo que hace el paquete y explicar como lo hace.

Como una particularidad, diré que el paquete esta escrito en lenguaje fortran, puede ser usado en la Burroughs 7800 y en las micros Printaform, Columbia, Corona PC o en las que se emplea el compilador profort (IBM).

2. REGRESION NO LINEAL.

Supongamos que los datos de respuesta Y observados correspondientes a los de entrada X son generados de acuerdo

al modelo:

$$Y = f(\beta) + e, \qquad (1)$$

donde $Y' = [Y_1, \ldots, Y_n]$ nx1

$f'(\beta) = [f(X_1, \beta), \ldots, f(X_n, \beta)]$ nx1

$\beta' = [\beta_1, \ldots, \beta_p]$ px1

$e' = [e_1, \ldots, e_n]$ nx1

Se supone que los errores se distribuyen como una normal con media cero y varianza σ^2.

El estimador de mínimos cuadrados del modelo (1), es el vector β que minimiza a:

$$S(\beta) = (Y - f(\beta))'(Y - f(\beta)). \qquad (2)$$

Para encontrar el estimador de mínimos cuadrados necesitamos diferenciar la ecuación 2 con respecto a β, las p ecuaciones normales que se obtienen de este proceso son también no lineales en β y se resuelven para β. Encontrar β resolviendo las ecuaciones normales no es fácil, para ello se proponen métodos iteractivos, uno de los métodos propuestos para tal fin es el denominado método modificado de Gauss-Newton propuesto por Hartley (1961).

Dentro del paquete existe una subrutina que contiene el método iteractivo de Gauss-Newton y uno de los argumentos de salida de esta subrutina contine el valor estimado de β.

Propiedades estadísticas de $\hat{\beta}$.

Bajo ciertas condiciones de regularidad $\hat{\beta}$ es un es-

timador consistente de β. Las condiciones de regularidad pueden encontrarse en los artículos de Jennerich (1969) y Burguete,Gallant,Souza(1983).

A nivel de resumen resaltamos estas propiedades:

$\hat{\beta}$ es un estimador consistente de β,

$\hat{\beta}$ converge casi seguramente a β,

$S^2=(1/(n-p))S(\beta)$ es un estimador consistente de σ^2 y

S^2 converge casi seguramente a σ^2.

$((1/n)F'(\beta)F(\beta))$ converge casi seguramente a la matriz Σ

$\sqrt{n}(\hat{\beta} - \beta)$ converge en distribución a una normal p-variada con media 0 y matriz de varianza covarianza σ^2 .

La importancia práctica de estas propiedades distribucionales es su uso para intervalos de confianza de los párametros desconocidos β . Y en pruebas de hipótesis de estos mismos párametros.

Con los resultados obtenidos de las propiedades estadísticas de β se puede obtener el valor:

$$t=\hat{\beta}/\sqrt{S^2\hat{C}_{ii}} \qquad (3)$$

donde $\hat{C}_{ii}$ son los elementos de la diagonal de la matriz $(F'(\hat{\beta})F(\hat{\beta}))^{-1}$.

Usando la distribución t de Student podemos encontrar intervalos de $(1 - \alpha)\%$ de confianza para el parámetro β. Con la expresión (3) podemos decidir rechazar o no Ho:$\beta = \beta$.

Con lo mencionado hasta aquí diremos que el paquete propuesto presenta un resumen estadístico de los resultados estimados esto es:

La matriz de varianza y covarianza asintótica de los parámetros β sea:

$$S^2 (F'(\beta)F(\beta))^{-1}$$

El coeficiente de variación $CV=\bar{Y}/S_y$

El coeficiente de determinación $r^2=1-S(\beta)/S(Y)$

Tamaño de la muestra n

Suma de cuadrados del error $S(\hat{\beta})$

La desviación estandar del error

La media $\bar{Y}$

El valor estimado β y el valor $t = \hat{\beta} / \sqrt{S^2 \hat{C}_{ii}}$.

Ejemplo: Para mostrar el funcionamiento y validez de los resultados obtenidos al usar el paquete que se propone en este trabajo, nos apoyamos en el artículo de Gallant (1975). Donde se usó el modelo:

$$Y_t = \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_4 \exp(\beta_3 X_{3t}) + e_t \qquad (4)$$

Los datos de entrada para este modelo corresponden a un diseño experimental con dos tratamientos (X_{1t}, X_{2t}) para un material experimental y la edad X_{3t} que afecta la respuesta exponencialmente.

En general la manera de entrar al paquete debe ser a través de generar un programa principal. En el deben incluirse la definición de los argumentos de la subrutina RREST, la que genera el resumen estadístico, los datos de respuesta y de entrada, en este caso Y_t, X_{1t}, X_{2t}, X_{3t}, el tamaño de la muestra, llamar la subrutina RREST, agregar la función y la subrutina correspondiente a la derivada parcial de la función que expresan al modelo con el cuál se generaron los datos.

El agregar la función y su derivada parcial en el programa se puede evitar, si el modelo que al usuario interese coincide con los ya integrados al paquete, estos son ciertos modelos de crecimiento y modelos sigmoidales. Para saber cuales son estas funciones, y además la documentación de las subrutinas que componen el paquete referirse a Domínguez (1986).

Prueba de hipótesis sobre β.

Por otro lado si uno desea probar alguna hipótesis sobre los parámetros del modelo. Por ejemplo:

$H0: h(\beta)=0 \quad$ vs $\quad H1: h(\beta)\neq 0$

donde $h: R^{p} \rightarrow R^{q}$. $H(\beta)=(\partial/\partial\beta')h(\beta)$.

Suponiendo que el rango H es q con q≤p.

Entonces el estadístico de prueba de Wald rechaza $H0: h(\beta)=0$ cuando:

$$S=h'(\hat{\beta})(\hat{H}'\hat{C}\hat{H}')^{-1}h(\hat{\beta})/qs^2$$

Excede el valor de $F(q,n-p)$, donde $H=H(\hat{\beta})$ y $C=(F'F)^{-1}$ Gallant (1975).

Ejemplo: Usando el modelo (4) y planteando la hipótesis $H0: h(\beta)=0$ vs $H1: h(\beta)\neq 0$, con $h(\beta)=\beta_1$. La información que tenemos del programa es la siguiente:

El valor del estadístico de Wald S=4.218050.

S se distribuye como una F(1,26).

En esta situación como las hipótesis que se formulen sobre los parámetros pueden ser muy variadas, aquí no se han agregado las funciones $h(\beta)$ ni sus derivadas. Por ello es necesario cuando se quiera hacer pruebas sobre los parámetros agregar las funciones $h(\beta)$ y las subrutinas que contengan las

derivadas parciales de $h(\beta)$.

En resumen el procedimiento para usar el paquete se puede ver de la siguiente forma:

```
                    Instrucciones Iniciales.
                    $RESET FREE
                    $SET AUTOBIND
                    $BIND=FROM(ISJD)OBJECT/MICROB
                    Definición de los argumentos.
Programa principal.-
                    Datos.
                    CALL nombre (argumentos)
                    STOP
                    END.
Función modelo      FUNCTION nombre (argumentos)
Subrutina derivada parcial del modelo
                    SUBROUTINE nombre(argumentos)
Función hipótesis parámetro
                    FUNCTION nombre (argumentos).
Subrutina derivada parcial de la función hipótesis.
                    SUBROUTINE nombre(argumentos).
```

En las siguientes dos hojas, se describe el programa principal para ajustar el modelo que se muestra en el ejemplo, enseguida se ilustran los resultados obtenidos al correr este programa, primero ajustado por mínimos cuadrados (ver hoja 9),luego la metodología robusta (hoja 10).

Para uasr este paquete en la Burroughs,primero se crea

```
$RESET FREE
$SET AUTOBIND
$BIND= FROM (ISJD) OBJECT/MICROB
FILE    5(KIND=REMOTE)
FILE    6(KIND=REMOTE,MAXRECSIZE=22)
       IMPLICIT REAL*8(A-H,O-Z)
       EXTERNAL F6,PARF6,SQUARE,HHIP,PARW
       REAL*8 THAT(4),TO6(4),TH6(4)
       REAL*8 Y(30),X1(30),X2(30),X3(30),Z(3,30)
C
C   VALORES INICIALES DE LOS PARAMETROS
C
        DATA TO6/-0.04866D0,1.038D0,-1.76D0,-0.5292D0/
C
C   CONJUNTO DE DATOS DE LA VARIABLE DEPENDIENTE
C
        DATA Y/.98610D0,.95482D0,1.02324D0,.96263D0,.98861D0,.98982D0
      1,.66768D0,.96822D0,.59759D0,1.01962D0,1.04255D0,.97526D0
      2,.80219D0,.95106D0,.50811D0,1.03848D0,1.04184D0,.90475D0
      3,1.05026D0,1.03437D0,1.01214D0,.55107D0,.98823D0,.99418D0
      4,.69163D0,1.04343D0,1.04969D0,1.01046D0,.97658D0,.91840D0/
C
C   CONJUNTO DE DATOS DE LA VARIABLE INDEPENDIENTE
C
        DATA X1/1.D0,1.D0,1.D0,1.D0,1.D0,1.D0,1.D0,1.D0,1.D0,1.D0
      1,1.D0,1.D0,1.D0,1.D0,1.D0,0.D0,0.D0,0.D0,0.D0,0.D0,0.D0,0.D0
      2,0.D0,0.D0,0.D0,0.D0,0.D0,0.D0,0.D0,0.D0,0.D0/
        DATA X2/1.D0,1.D0,1.D0,1.D0,1.D0,1.D0,1.D0,1.D0,1.D0,1.D0
      2,1.D0,1.D0,1.D0,1.D0,1.D0,1.D0,1.D0,1.D0,1.D0,1.D0,1.D0,1.D0
      3,1.D0,1.D0,1.D0,1.D0,1.D0,1.D0,1.D0,1.D0,1.D0/
        DATA X3/6.28D0,9.11D0,8.11D0,6.58D0,6.52D0,9.86D0,.47D0
      1,4.07D0,.17D0,4.39D0,4.73D0,8.90D0,.77D0,4.51D0,.08D0
      2,9.86D0,8.43D0,1.82D0,5.02D0,3.75D0,7.31D0,.07D0,4.61D0
      3,6.99D0,.39D0,9.42D0,3.02D0,3.31D0,2.65D0,6.11D0/
C
C  CONDICIONES
C
        N=30
        E1=1.D-4
        E2=1.D-4
        LIMIT=40
        TAO=1
C
C
C
        DO 25 J=1,N
        Z(1,J)=X1(J)
        Z(2,J)=X2(J)
        Z(3,J)=X3(J)
 25     CONTINUE
C
C       CALCULA EL ESTADISTICO DE WALD
C
```

```
5310          CALL RREST(TO6,4,Z,3,N,Y,F6,PARF6,TH6,THAT,1.D0,FLMODE,1)
5400          CALL RRWALD(1.D0,4,Z,3,N,Y,F6,PARF6,TH6,THAT,HHIP,PARW,1,1
5600 C
5700 C        DESCRIBE DATOS ESTADISTICOS
5800 C
5900          ELMODE='PRUEBA'
6100          CALL RREST(TO6,4,Z,3,N,Y,F6,PARF6,TH6,THAT,TAO,ELMODE,0)
6110          CALL RRWALD(TAO,4,Z,3,N,Y,F6,PARF6,TH6,THAT,HHIP,PARW,1,0)
6200          STOP
6300          END
6400          FUNCTION F6(X,T,K,IP)
6500          IMPLICIT REAL*8(A-H,O-Z)
6600          REAL*8 X(K),T(IP)
6700          F6=T(1)*X(1)+T(2)*X(2)+T(4)*DEXP(T(3)*X(3))
6800          RETURN
6900          END
7000          SUBROUTINE PARF6(DPF6,IP,X,K,T)
7100          IMPLICIT REAL*8(A-H,O-Z)
7200          REAL*8 X(K),T(IP),DPF6(IP)
7300          DPF6(1)=X(1)
7400          DPF6(2)=X(2)
7500          DPF6(3)=T(4)*X(3)*DEXP(T(3)*X(3))
7600          DPF6(4)=DEXP(T(3)*X(3))
7700          RETURN
7800          END
```

REGRESION MINIMOS CUADRADOS

** PARA EL MODELO :**

*MATRIZ DE VAR-COV ASINTOTICA DE LOS PARAMETROS**

.159690E-03	-.788774E-04	-.177478E-03	-.441848E-04
-.788774E-04	.989645E-04	.608335E-03	-.185612E-05
-.177478E-03	.608335E-03	.268092E-01	.236150E-02
-.441848E-04	-.185612E-05	.236150E-02	.659668E-03

** COEF. DE VARIACION **	0.17981
** COEF. DE DETERMINACION **	0.99998
** TAMA@O DE LA MUESTRA **	30
SUMA DE CUADRADOS DEL ERROR	0.030558
**DESV ESTANDAR DEL ERROR **	0.034283
** MEDIA **	0.924767

PARAMETRO	**T-STUDENT**
T(1)=-.259534107486E-01	-2.053789
T(2)= .101567132996E+01	102.097120
T(3)=-.111572890733E+01	-6.814235
T(4)=-.504843599533E+00	-19.655955

ESTIMACION MINIMOS CUADRADOS

** EL VALOR DEL ESTADISTICO DE WALD ** S= 4.218050

** S SE DISTRIBUYE COMO UNA F(1,26) **

```
        * REGRESION M-ROBUSTA **

             ** PARA EL MODELO :PRUEBA**

*MATRIZ DE VAR-COV ASINTOTICA DE LOS PARAMETROS**


 .131592E-03 -.651352E-04 -.149856E-03 -.366365E-04
-.651352E-04  .819238E-04  .504304E-03 -.130680E-05
-.149856E-03  .504304E-03  .216597E-01  .190880E-02
-.366365E-04 -.130680E-05  .190880E-02  .538988E-03


        ** COEF. DE VARIACION VAR.DEP.**          0.17981

        **      TAMA@O DE LA MUESTRA     **        30

        ** M-EST. PARAMETRO DE ESCALA  **         0.034950

        **       MEDIA VAR. DEP.         **       0.924767


        **PARAMETRO**                        **VALOR DE T**

        T(1)=-.285015347070E-01               -2.484587

        T(2)= .101795353977E+01              112.466403

        T(3)=-.110621725806E+01               -7.516476

        T(4)=-.504860188343E+00              -21.746103


        **      ESTIMACION ROBUSTA      **


        ** EL VALOR DEL ESTADISTICO DE WALD ** S=     6.173170


        ** S SE DISTRIBUYE COMO UNA JI CUADRADA CON       ,GL= 1 **
```

el programa principal con funciones y subrutinas de interes, junto con las instrucciones iniciales descritas anteriormente en el resumen, se compila (compile) el programa y luego se corre (run).

El uso de este paquete en las micros es análogo, primero se crea el programa principal através de un editor (sin usar las instrucciones iniciales), despues se compila usando el profort, enseguida se encadena con un linker los siguientes archivos el que contiene el programa principal + MINCUA + RESTAD + ESTROB*, con esto se crea un archivo cuyo nombre es el dado al programa principal con la extención .EXE, finalmente este último se corre, obteniendo los resultados que aparecen en las hojas 9 y 10.

3.Estadísticos para probar hipótesis no anidadas..

En el quehacer científico se llega a dar la situación de que diferentes teorías al tratar de explicar un mismo fenómeno dan lugar a modelos distintos que tiene la característica de no poderse representar uno como caso particular del otro y viceversa. A este tipo de modelos se les denominará no anidados.

Estos modelos se pueden expresar por:

$$H_0: Y=f(X,\beta)+e_0$$

$$H_1: Y=g(X,\alpha)+e_1.$$

H0 y Hi son no anidados si no existe una transformación $\delta=h_0(\tau)$ tal que $f(x,h_0(\tau))=g(x,\tau)$ para cada (x,τ), $\forall_{\tau\in}$, ni existe $h_1(\delta)$ tal que $f(x,\delta)=g(x,h_1(\delta))$ para cada (x,δ) $\forall\,\delta\in\Delta$ ..

Para probar hipótesis no anidadas se han propuesto varios estadísticos entre ellos podemos citar los Pesaran y Deaton (1978) (CPD), los obtenidos por los procedimientos J y P de Davidson y Mac Kinnon (1981) JDM, PDM, los obtenidos por la metodología de Aguirre-Gallant (1983). Un estudio comparativo de las propiedades de estos estadísticos se presenta en Domínguez (1986)..

En este paquete se han programado subrutinas para cada uno de los estadísticos, obteniendo en ellos el valor respectivo del estadístico de prueba, considerando diferentes parejas de modelos. Cabe observar que en el cálculo de los valores de los estadísticos de Resaran y Deaton y en los procedimientos J y P de Davidson y Mac Kinnon los errores en los modelos se suponen se distribuyen como una normal con media cero y varianza σ^2.

Como reporte de salida se tiene

Pesaran y Deaton:

Ho: Nombre del modelo.

Hi: Nombre del modelo.

el valor del estadístico

Davidson y Mac Kinnon:

Ho: Nombre del modelo.

Hi: Nombre del modelo.

el valor del estadístico.

Aguirre-Gallant:

Ho: Nombre del modelo.

Hi: Nombre del modelo.

el valor del estadístico.

REFERENCIAS.

Aguirre-Torres. V (1984). Testing several non-nested regres sions simultaneously. A nonparametric approach. Comu nicaciones tècnicas, serie naranja. No.372 IIMAS, UNAM, MEXICO.

Aguirre-Torres. V, Gallant A.R. (1982) On choosing between two nonlinear models estimated robustly.Unpublished Manuscript.

Burguete,F.,Gallant,R.y Souza,G.(1983).On unification of the asymptotic theory of nonlinear econometric models. Econometric Review

Dominguez,J.(1986)Comparaciòn de pruebas para hipòtesis no anidadas en regresiòn no lineal.T.de maestria UNAM.

Davidson,R. and Mac Kinnon, J.(mayo 1981) Several tests for model especifications in the presence of alternative hipotheses. Econometrica, Vol.49 No.3

Gallant,A.R.(1975). Nolinear Regresion. The American Statis tican,Vol 29,No 2.

Jennrich,R.I. (1969). Asynmptotic properties of non-linear least squares estimators. The Annals of Mathematical Statistics. Vol.40, No 2, 633-643.

Pesaran,A.H.,Deaton,A.S.(1978)."Testing non-nested nonlinear regression models."Econometrica, 46,677-694.

Ryan-McFarland Corporation (1984).Installation and use.IBM personal computer professional fortran.

UNA APLICACION DEL ESCALAMIENTO MULTIDIMENSIONAL EN UN ESTUDIO EN PERINATOLOGIA

Guillermina Eslava

IIMAS-UNAM

I. INTRODUCCION

En un estudio realizado por investigadoras del Instituto Nacional de Perinatología, se evaluaron a través de un conjunto de reactivos a un grupo de neonatos clasificados en cuatro subgrupos.

Los datos registrados en esta investigación son referidos dentro de la Estadística con el nombre de "Datos Categóricos Multivariados".

Con el conjunto de mediciones se desea construir variables numéricas con el fin de hacer un análisis exploratorio y descriptivo de la información.

En cada neonato se midieron 74 reactivos. El grupo total de neonatos se clasificó, de acuerdo al diagnóstico médico al momento del nacimiento, en cuatro grupos: SANOS, HIPERBILIRRUBINEMICOS, SANOS, de alto riesgo TERMINO y de alto riesgo PRETERMINO (período gestacional menor a nueve meses).

Debido al tipo de datos se usó el escalamiento multidimensional,

como una herramienta útil para un análisis exploratorio y descriptivo.

2. METODOLOGIA DE ANALISIS

El escalamiento multidimensional es un método de representación espacial en $\mathbb{R}^k$ de un conjunto de N objetos, individuos o características. La representación se caracteriza por reflejar en sus distancias interpuntuales la estructura cualitativa de parecido o similaridad entre individuos.

Se supone que se cuenta con medidas de similaridad o parecido entre individuos, características u objetos, estas medidas se pueden construir a partir de información disponible o pueden ser proporcionadas directamente por el investigador conocedor del problema. Se denota por:

$s(i,j)$ = similaridad entre individuo i y j.

$d(P_i,P_j)$ = distancia entre representante i y j en el plano en $\mathbb{R}^k$.

Por la forma en que se relacionan las distancias interpuntuales y las medidas de similaridad entre individuos, la metodología recibe diversos nombres. Matemáticamente el problema se reduce a un problema de minimización de una función de varias variables, es decir:

$$\underset{\{P_i,f\}}{\text{Min}}\ S = \frac{\sum_{i,j=1}^{N} (d(P_i,P_j) - f(s(i,j))^2}{\sum_{i,j=1}^{N} d^2(P_i,P_j)}$$

La solución corresponde a la configuración de puntos $\{P_i\}$ en un espacio $\mathbb{R}^k$ y una función f en un espacio determinado, tales que minimicen la función S, ésta función es conocida, en la literatura estadística de escalamiento multidimensional, como "stress".

Se distinguen básicamente tres tipos de escalamiento multidimensional y la forma numérica de solución es diferente:

- Cuando el espacio de puntos sobre el cual se busca, el mínimo es $\mathbb{R}^N$, donde N es el número de sujetos a representar y la función f es

$$d(i,j) = \{2 - 2S(i,j)\}^{1/2} = f(s(i,j))$$

este caso se conoce también con el nombre de "Coordenadas Principales". Este este caso es posible enocntrar una configuración solución donde la función stress alcanza el valor mínimo cero, sin embargo el espacio de representación es de dimensión N-1.

- Un enfoque más amplio que el anterior es considerar un espacio de dimensión K menor o igual a N-1, y un espacio de funciones paramétricas. Se minimiza la función stress sobre estos dos espacios. De esta forma es posible encontrar representaciones en un espacio de dimensión menor a N-1. Este enfoque de solución es conocido con el nombre de escalamiento multidimensional métrico.

- El caso más general, recomendable cuando los anteriores no proporcionan representaciones satisfactorias es buscar el mínimo de la función stress sobre el espacio con dimensión K menor a N-1 $\mathbb{R}^k$ y sobre el espacio de funciones monótonas crecientes (no paramétricas).

  Al considerar que la función liga f sea monótona creciente, permite captar relaciones lineales y no lineales entre similaridades y distancias permitiendo obtener una configuración solución en un espacio de dimensión menor que la que se obtiene con el procedimiento métrico.

3. APLICACION DEL ESCALAMIENTO NO METRICO AL ANALISIS DE LOS DATOS

El primer paso en el análisis es el de formar las matrices de similaridad en base a la información que se tiene. Para esto hacemos explícita la información disponible.

EDAD (días)	01	10	15	20	30
Sanos	29	29	--	29	29
Hiperbilirrulinémicos	12	12	--	12	12
Término	16	--	27	--	30
Pretérmino	14	--	29	--	32
Total	81	97		97	103

Las flechas indican que los neonatos de 15 días se tomaron en los grupos de 10 y 20 días.

Para cada neonato se tienen 6 grupos de variables (reactivos) y en cada grupo hay varias mediciones. En detalle los grupos son:

Habituación	4 variables
Orientación	8 variables
Regulación de estado	4 variables
Rango de estado	5 variables
Regulación antonómica	3 variables
Funcionamiento motor	5 variables

Las matrices de similaridad se construyeron en base a los grupos de variables y se tiene una matriz por cada grupo de edad.

Denotándo por Kg el número de variables en el grupo g de reactivos, se define la similaridad entre los individuos i,j como

$$s(i,j) = \sum_{t=1}^{kg} s_{ijt}/Kg$$

donde $s(i_j t) = 1 - |X(i,t) - X(j,t)|/R(t)$ es el coeficiente de similaridad entre el individuo i y el j con respecto a la variable t. Las X(i,t) denotan el valor de la variable t en el neonato i y R(t) es el rango de variación teórico de la variable t.

Con este procedimiento se obtienen 20 matrices que resumen la información por edad y por grupo de variables. Estas son de diferentes dimensiones ya que se tiene diferente número de individuos a distintas edades.

Se tienen así las siguientes matrices de similaridad:

Kg	Grupo de reactivos	01	10 ó 15	15 ó 20	30
8	Orientacción	81	97	97	103
4	Reg. Estado	"	"	"	"
5	Rango	"	"	"	"
3	Reg Anton	"	"	"	"
5	Motor	"	"	"	"

Para condensar la información de cada matriz se efectuó un análisis de escalamiennto clásico. Es decir se buscaron componentes principales con la esperanza de tener uno o dos valores propios que explicaran un gran porcentaje de variación. Sin embargo los datos no reflejaron un componente dominante tal como se muestra en la Tabla.

Valor Propio	% total de la varianza explicada en el grupo de variables				
	Orientación	Reg Estado	Rango	Reg Anton	Motor
1	33	24	40	28	18
2	39	36	50	43	34
3	45	48	57	57	54

De esta Tabla se ve que con tres valores propios se explica a lo más el 57% de la variabilidad total y esto indica que hay que utilizar otra representación para reducir la dimensionalidad del espacio.

Para reducir la dimensionalidad se utilizó el escalamiento no métrico que se describió al inicio de la segunda sesión.

4. RESULTADOS DEL ESCALAMIENTO NO METRICO

Las configuraciones buscadas por grupo de variables fueron en una y dos dimensiones para poder hacer un análisis en la recta o en el plano y poder también estudiar el poder discriminatorio de parejas de grupos.

El stress se minimizó en dos pasos, en el primero se fijó la configuración de puntos y mediante una regresión monótona se encontró la función liga f, el siguiente paso minimizaba sobre la configuración.

El criterio para terminar la minimización fue el de obtener un stress bajo y una configuración interpretable.

De cada grupo de variables se obtuvo una variable unidimensional A para cada grupo de edad. Para una día de edad se obtuvo también una representación en dos dimensiones.

El stress asociado fue satisfactorio y bajó al aumentar el número de variables para la representación de los N individuos. Por otro lado no es práctico tener configuraciones en dimensión alta ya que no son visualizables.

Se hizo finalmente un análisis por parejas graficando los individuos como puntos en los planos A_i, A_j; para todas las parejas de variables. Esto con el fin de estudiar si parejas de grupos de reactivos ayudaban a discriminar mejor que un sólo grupo.

Los resultados obtenidos de este análisis pueden resumirse así:

- La variable representante de Orientacción no discrimina a los diferentes grupos. Por otra parte aparecen dos individuos que se alejan del grupo total y que deben de ser estudiados con más cuidado.

- La variable Rango no tiene poder discriminatorio.

- Las variables Reg Estado y Reg Anton separan a los neonatos clasificados médicamente como Término y Pretérmino.

- Las variables Motor con Reg Estado separan a Sanos de Pretérmino. Los Hiperbilirrubinémicos aparecen mezclados con los sanos.

5. CONCLUSIONES

Desde el punto de vista de metodología debe de usarse el escalamiento no métrico cuando la estructura de los datos es tal que los valores propios de la matriz de similaridad sean muy parecidos.

En casos de valores propios muy diferentes el escalamiento métrico da configuraciones satisfactorias y es menos costoso computacionalmente.

Desde el punto de vista del diagnóstico clínico se observa:

- Los grupos de variables que mejor discriminan son, de acuerdo con la edad, Motor, Reg Estado, Reg Antonómica.

- El grupo Orientación no discrimina entre grupos neonatos.

Para completar el estudio se sugiere considerar mayor información para los Hiperbilirrulinémicos con el fin de poder identificarlos ya que en este análisis se mezclan los individuos sanos.

REFERENCIAS

ESLAVA, Guillermina "Análisis Multivariado de Datos Categóricos" Tesis de Maestría. Septiembre 1986.

Estructura comunitaria de la fauna crustacea decapoda en la plataforma continental del noreste del Golfo de Mexico.

ACT. ALEJANDRO GONZALEZ RULLAN

Egresado de la Maestría en Ciencias del Mar.

Introduccion.

Desde el punto de vista ecologico una comunidad es una poblacion formada por individuos de diferentes especies, que viven en un espacio continuo, delimitado de manera convencional. La forma en que se distribuyen los recursos entre las especies de una comunidad asi como los patrones de distribucion espacial y temporal constituyen la estructura de la comunidad.

En el medio ambiente marino el estudio de las comunidades bentonicas, es decir las que habitan dentro o sobre el fondo del mar, es particularmente importante porque en ocasiones algunos de sus integrantes son explotados comercialmente o por su relacion con otras especies de peces de importancia comercial. Ademas, son estas comunidades las que resienten con mayor fuerza la contaminacion producida por los derrames de petroleo.

El presente estudio trata la estructura de una de las comunidades bentonicas, la de los crustaceos decapodos, es decir la compuesta por especies comunmente conocidas como camarones, cangrejos, etc. En particular trata de determinar patrones en la distribucion de los organismos y las relaciones de aquellos con la temperatura, profundidad, tipo de sedimento y epoca del año, mediante metodos multivariados de agrupacion y ordenacion. El area de estudio abarca la plataforma continental, que es la parte del fondo del oceano donde la pendiente es menos pronunciada, desde Cabo St. George hasta la bahia de Mobile. Desde el punto de vista ecologico y zoogeografico, la existencia de fenomenos fisicos como la descarga del Mississippi, la proximidad del Cañon de De Soto y las diferentes masas de agua de este sector del Golfo de Mexico, como son la corriente de Lazo, el giro estuarino de Florida y el giro de la Bahia de Florida influyen fuertemente sobre la complejidad en la distribucion de los organismos que se observa en el area y en la demarcacion de fronteras biogeograficas.

Material y Metodo.

La informacion se obtuvo del estudio realizado por Soto (1972) del material colectado durante la operacion exploratoria de arrastres,. llevada a cabo a bordo de un barco de la Universidad Estatal de Florida, durante el

periodo de octubre de 1970 a octubre de 1971. Esta consistio de 108 arrastres que se obtuvieron en dos tipos de transectos: el primero fue uno regular frente a la bahia Apalachicola, desde los 18 m. hasta el borde de la plataforma contnental. El segundo fue mas variable y comprendio desde el area frente a Cabo San Blas hasta la seccion oriental del Delta del Mississippi. Ambos fueron trazados siguiendo el contorno de las lineas isobatas en zonas donde fuera posible realizar arrastres. Se colectaron especimenes de 120 especies.

Para poder estudiar la variacion estacional los arrastres se agruparon de acuerdo a la epoca del año en que fueron realizados y posteriormente se procedio a la reduccion del numero de datos, para cada estacion se eliminaron las especies que aparecieron en menos de tres localidades y las colectas que contenian menos de dos de las especies restantes. Estas reducciones son necesarias porque las especies con una o dos apariciones generalmente carecen de un patron de distribucion y por otra parte, las colectas con una sola especie pueden causar confusion en la interpretacion de los resultados del analisis de cumulos, ademas de no contener una captura representativa de la fauna decapoda.

El numero de arrastres en cada estacion del año no fue el mismo para todas, hubo una diferencia en la distribucion de los mismos a lo largo de la plataforma continental, como puede verse en las figuras 1 a 4. Ademas, la distribucion de los arrastres en las diferentes profundidades tampoco fue regular. Se pueden distinguir tres zonas, de acuerdo con la intensidad del muestreo: la primera abarca la plataforma continental al occidente del Cañon de De Soto, la segunda se localiza frente a las bahias Choctawhatchee y St. Andrew y la tercera frente a los cabos San Blas y St. George, esta ultima es la que presenta mas regularidad en el muestreo.

Para estudiar la variacion en la distribucion de las especies que aparecieron mas frecuentemente, tanto en el espacio como en el tiempo, se utilizo un indice de abundancia (Musick y McEachran, 1972) relativa y se elaboro una grafica con la epoca del año, profundidad y zona de la plataforma.

Para reconocer grupos de especies, inicialmente se utilizo el analisis de grupos recurrentes, para datos de ausencias y presencias, tomando como indice de afinidad a la media geometrica de la proporcion de apariciones simultaneas, corregida por el tamaño de la muestra, el cual fue formulado por Fager y McGowan (1963).

Posteriormente se usaron el analisis de cumulos y el de correspondencias, ya que estos metodos han sido utilizados con exito para estudiar la estructura de las comunidades bentonicas. En el analisis de cmulos se utilizo el coeficiente de disimilitud de Camberra (Lance y Williams,

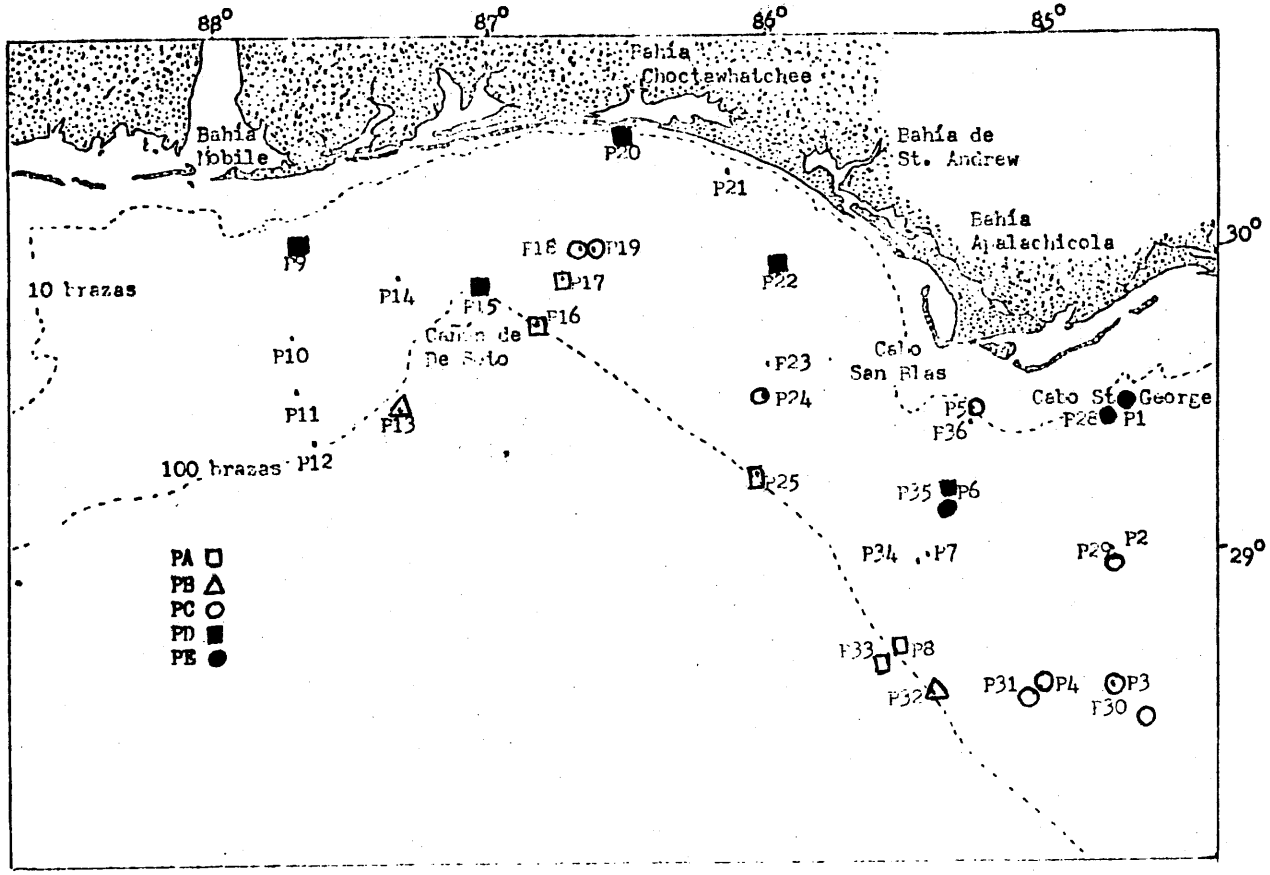


Fig. 1. Localización de los arrastres realizados durante la primavera.

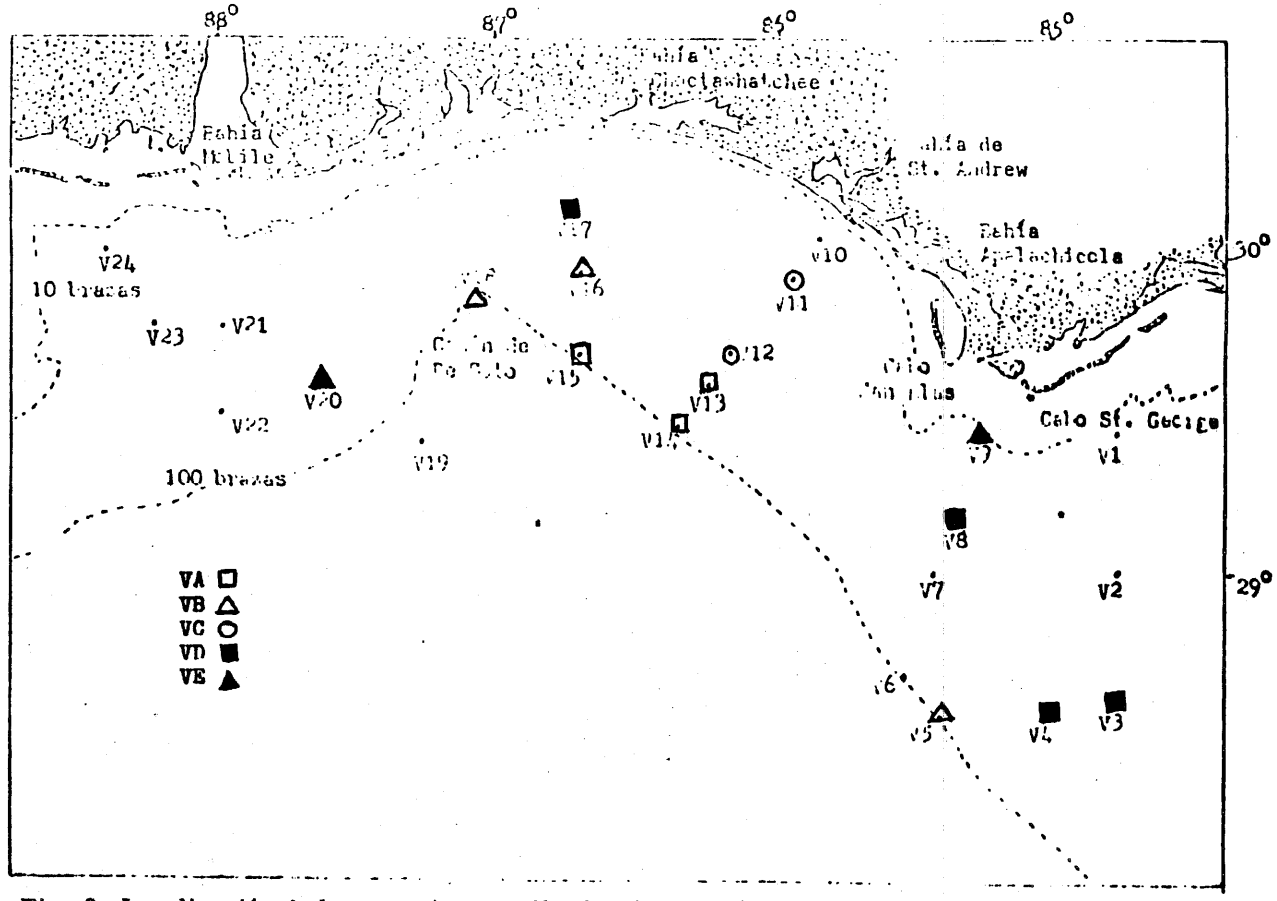


Fig. 2. Localización de los arrastres realizados durante el verano.

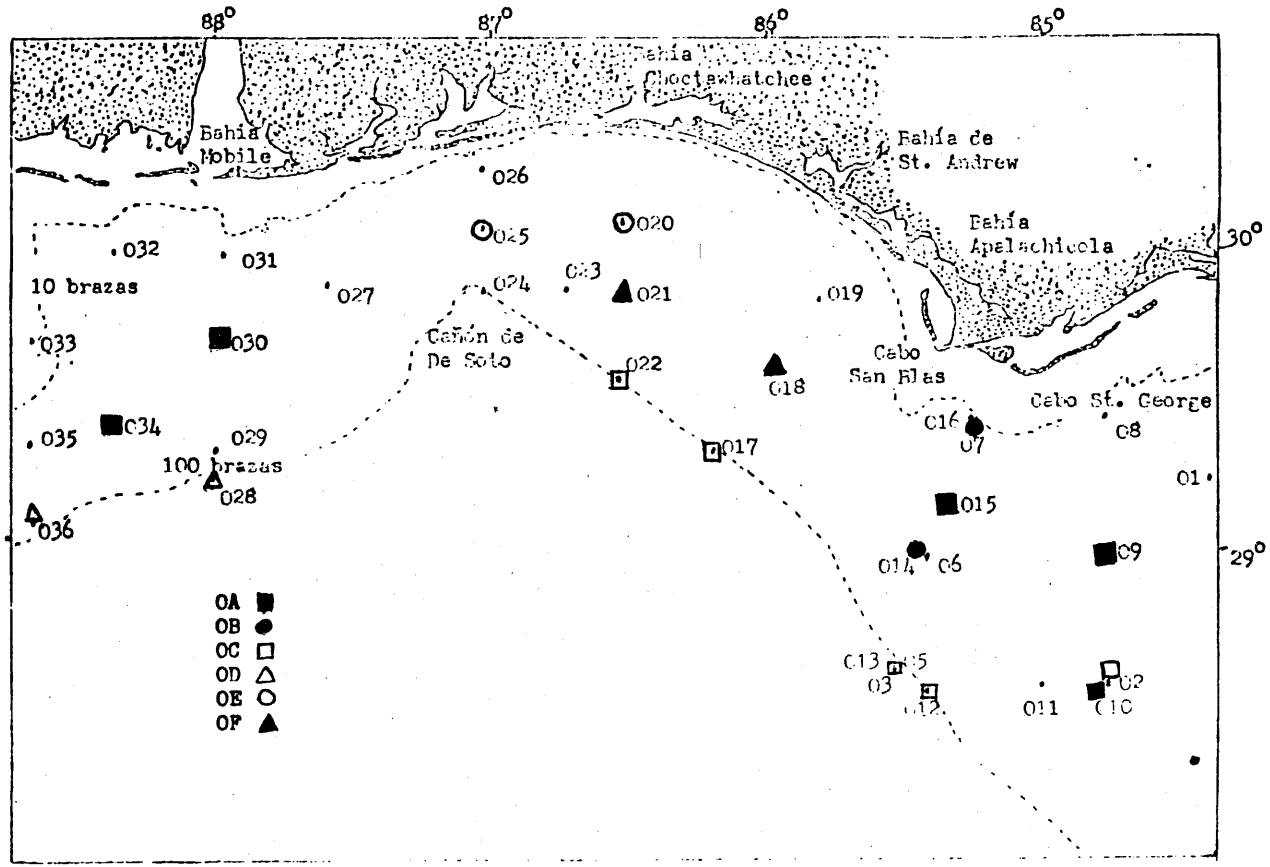


Fig. 3. Localización de los arrastres realizados durante el otoño.

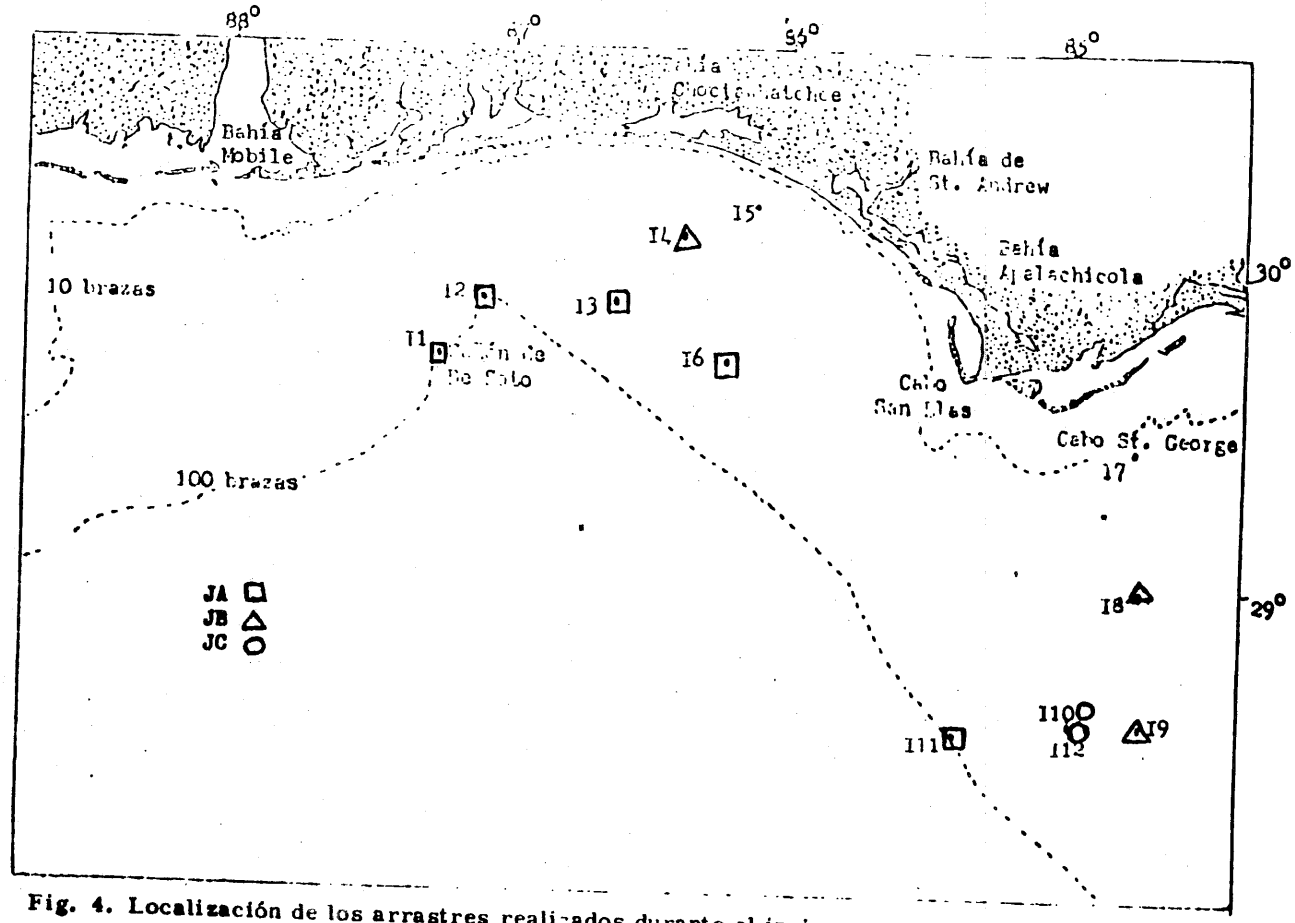


Fig. 4. Localización de los arrastres realizados durante el invierno.

1967) ya que resulta efectivo cuando los organismos tienen habitos gregarios, como la mayoria de los organismos bentonicos y porque da a las especies raras tanto peso como a las abundantes. Previamente se utilizo una estandarizacion doble simultanea para eliminar problemas de escala (Boesch, 1973). Se escogio la estrategia de agrupamiento jerarquico aglomerativo de la distancia media y se elaboraron los dendrogramas correspondientes para el analisis normal e inverso (Reyes, 1978). El analisis se aplico a las colectas de cada estacion y a las de todo el año juntas. Para poder estudiar la relacion entre los grupos de especies y de arrastres se calcularon indices de constancia y fidelidad.

Para el analisis de correspondencias (Malmgrem et al., 1978) se utilizo una transformacion logaritmica de los datos con el fin de reducir las diferencias escalares. Este metodo es el mas util de los metodos multivariados de ordenacion para detectar patrones de interrelacion de especies y colectas y para detectar cambios a lo largo de un gradiente, es decir siguiendo el cambio continuo de alguna variable.

Los indices de diversidad de la teoria de la informacion expresan la incertidumbre en la prediccion de la identidad de una especie escogida al azar de una coleccion. Con el fin de comparar las diferentes localidades se calcularon indices de diversidad, riqueza de especies (Margalef,1958) y equitatividad (Pielou, 1966) de cada una de las 108 colectas, tomando en cuenta a todas las especies.

Resultados.

En la tabla 1 se presentan los numeros de especies y colectas por epoca del año, antes y despues de la reduccion de datos. El numero de arrastres que se realizaron durante el invierno fue considerablemente inferior que el de los realizados en las otras estaciones, por lo que se considero conveniente realizar tambien el analisis para las especies que aparecieron al menos en dos ocasiones en el invierno. En el analisis global se incluyeron las especies que aparecieron en al menos tres ocasiones en alguna estacion del año.

Abundancia relativa de las especies numericamente dominantes.

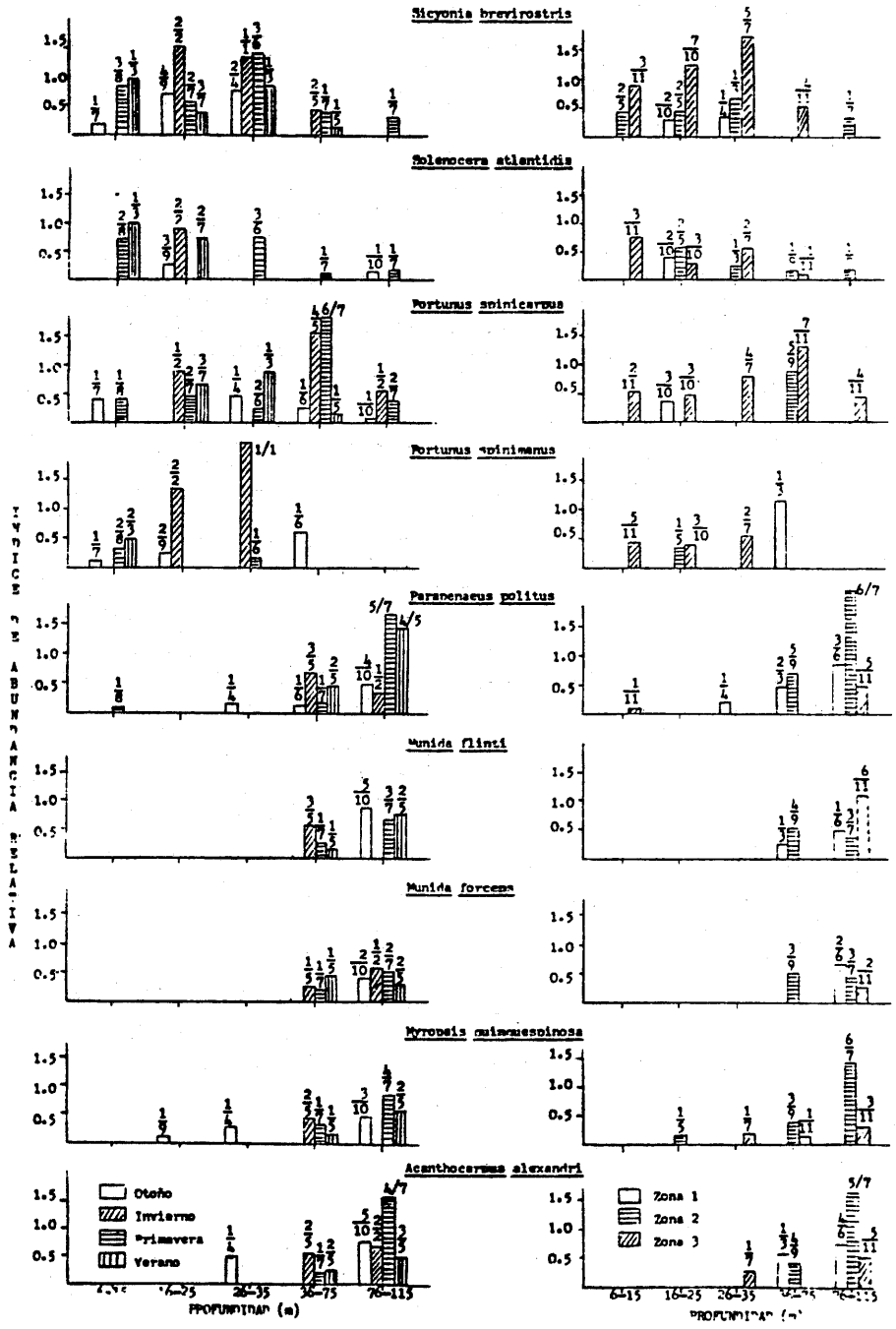
En la figura 5 se representa el indice de abundancia relativa de las especies que mas frecuentemente se capturaron a lo largo de todo el año, de acuerdo a la profundidad y a la estacion del año o a la zona de la plataforma. Las primeras 4 aparecieron mas frecuentemente en las partes interna y media de la plataforma continental, mientras que las 5 restantes aparecieron fundamentalmente en el borde externo de la plataforma. La mayoria de las especies

Tabla 1. Numero de especies y de arrastres antes y despues de la reduccion de datos

Periodo.	No. de arrastres realizados.	No. de arrastres despues de la reduccion.	No. de especies que aparecieron.	No. de especies despues de la reduccion.
primavera	36	26	80	30
verano	24	16	73	20
otoño	36	18	69	19
invierno (3)*	12	10	31	10
invierno (2)**	12	11	31	18
total	108	76	120	42

* Tomando en cuenta las especies que aparecieron al menos 3 veces.

**Tomando en cuenta las especies que aparecieron al menos 2 veces.


Sicyonia brevirostris
Solenocera atlantidis
Portunus spinicarpus
Portunus spinimanus
Parapenaeus politus
Munida flinti
Munida forceps
Myropsis quinquespinosa
Acanthocarpus alexandri
INDICE DE ABUNDANCIA RELATIVA
Otoño
Invierno
Primavera
Verano
Zona 1
Zona 2
Zona 3
PROFUNDIDAD (m)

restantes tambien mostraron patrones de distribucion restringidos a la partes interna y media o externa de la plataforma continental y aparentemente su distribucion esta mas influenciada por la presencia del Cañon de De Soto, la descarga del rio Mississippi y las corrientes, ya que aparecen solamente al orente del Cañon de De Soto.

Grupos recurentes.

El analisis de grupos recurrentes s realizo utilizando el valor de 0.5 como valor critico para determinar la existencia de afiidad entre las especies. Solamente se formaron dos grupos con dos especies cada uno, por lo que no resulto satisfactoria la clasificacion.

Analisis de cumulos.

Analisis normal.

Los dendrogramas resultantes del analisis normal para cada estacion del año se truncaron primero para reconocer dos grupos. Se encontro que uno esta formado por arrastres realizados entre 20 y 100 metros de profundidad y el otro por los obtenidos a mas de 100 metros, excepto en el del otoño que incluye un grupo formado por colectas obtenidas a mas de 100 metros y por 3 a profundidades menores.

En las figuras 1 a 4 se puede observar la localizacion de los sitios donde se realizaron los arrastres en cada estacion del año asi como su clasificacion al truncar los dendrogramas para obtener dos grupos en la parte externa de la plataforma continental y tres en las partes interna y media.

Para el analisis de todos los arrastres juntos se consideraron 42 especies y 76 arrastres; debido a esto se trato de que la clasificacion tuviera mas grupos. Casi todos los grupos incluyen arrastres realizados en varias estaciones del año (tabla 2). Las diferencias en el numero de arrastres de cada epoca del año pueden ser debidas a la variacion en la intensidad del muestreo, particularmente frente a Cabo St. George. Las diferencias en los grupos con menos de tres arrastres pueden ser azarosas, por lo que la agrupacion parece depender de factores independientes de la epoca del año, como la profundidad y el tipo de sustrato. La distribucion de algunos grupos generados en el analisis global puede observarse en las figuras 6 y 7.

La composicion de los sedimentos en el area de estudio varia de acuerdo con la profundidad y la proximidad al Delta del Mississippi. Esto influye de manera general en la distribucion de los crustaceos decapodos al igual que en la de otros organismos bentonicos, con una reduccion de especies hacia en occidente del Cañon de De Soto. Sin

Tabla 2. Numero de arrastres por grupo, del analisis de cumulos normal global, y epoca del año.

No. de arrastres.	Grupos GA	GB	GC	GD	GE	GF	GG	GH	GI	GJ	GK	GL	GM	GN	GO	GP	GQ	GR
prim.-	3-	2-	3-	2-	1-	1-	2-	1-	0-	1-	0-	1-	2-	0-	3-	0-	1-	1
ver. -	1-	1-	1-	2-	1-	2-	0-	0-	2-	0-	1-	1-	2-	2-	2-	0-	0-	0
otoño-	1-	2-	1-	2-	1-	2-	0-	1-	0-	2-	0-	0-	2-	1-	0-	2-	1-	1
inv. -	1-	2-	1-	1-	0-	0-	0-	0-	0-	0-	1-	0-	2-	3-	0-	0-	0-	0

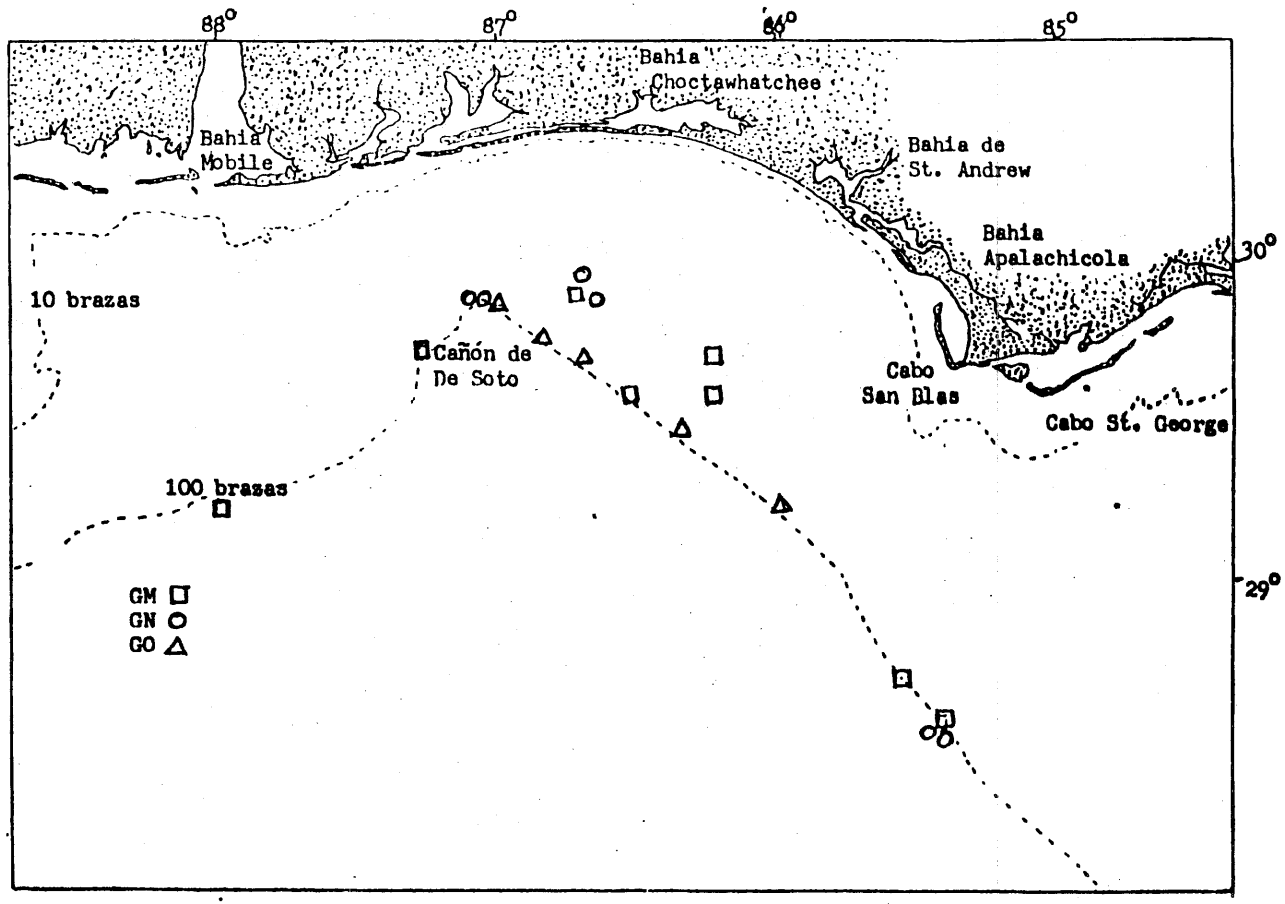


Fig. 6. Distribucion de los grupos generados en el analisis global.

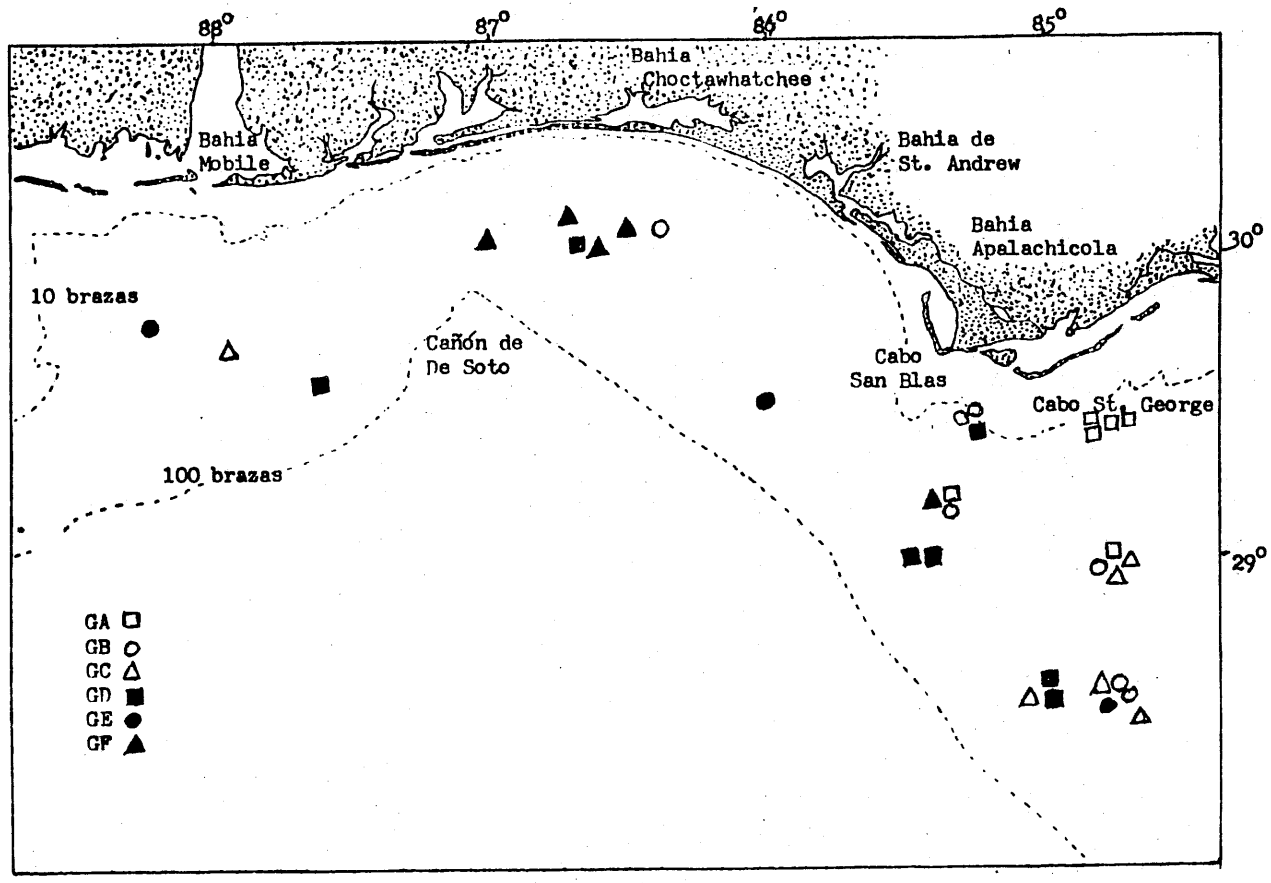


Fig. 7 Distribucion de los grupos generados en el analisis global.

embargo, esto no parece influir de modo definitivo en la formacion de grupos en el analisis de cumulos que en alguna ocasiones incluyen localidades de ambos lados del cañon, especialmente en la parte externa de la plataforma continental.

Analisis inverso.

Los dendrogramas del analisis de cumulos inverso se truncaron en el nivel en el que se reconocian mas de dos grupos en la parte externa de la plataforma continental y en las partes interna y media, tratando de que se agruparan la mayoria de las especies, pero tambien de que las especies tuvieran patrones de distribucion similares. Este proposito se logro mediante la revision de la tabla de especies y colectas. Debido al bajo numero de especies que aparecieron al menos tres veces en el invierno solamente se reconocio un grupo en las partes interna y media de la plataforma continental.

La mayoria de los grupos estan formados por especies que fueron capturadas a lo largo de casi todo el año, junto con especies que solo se presentaron en tres o mas arrastres en dos estaciones del año. Los grupos que solo incluyen organismos con distribucion temporal restringida, generalmente consistieron de dos especies.

Como en el analisis normal, en el analisis global no se observa una marcada agrupacion debida a diferencias estacionales.

En la tabla 3 se puede observar la correspondencia entre los grupos de arrastres y de especies, quedando estos ultimos restringidos casi exclusivamente al ambiente de la parte externa o al de las partes interna y media de la plataforma continental, por lo que en general se pueden reconocer dos componentes faunisticas. La primera, que habita en las partes interna y media de la plataforma continental,esta constituida por un mayor numero de especies y presenta mas fluctuaciones en su composicion a lo largo del año que la segunda, que habita la parte externa de la plataforma, posiblemente debido a que esta esta constituida por especies que habitan principalmente en la parte superior del talud. La presencia de un cambio abrupto en la composicion de las comunidades bentonicas ha sido reportado en varios estudios y puede decirse que en este caso ocurre entre los 110 y 120 metros de profundidad, aunque puede variar a lo largo de la plataforma. Esta no es la unica zona donde ocurre el cambio en la composicion de las comunidades de la plataforma continental, cerca de la zona costera, a profundidades cercanas a los 15 metros tambien se puede presentar, sin embargo en el presente estudio no se pudo detectar por la falta de un buen numero de muestras en esta zona.

Tabla 3. Valores de constancia de los grupos de especies en los grupos de arrastres, obtenidos en el analisis de cumulos usando la distancia media, para cada estacion del año y todo el año (global).

Primavera

Grupos de especies	Grupos de arrastres PA	PB	PC	PD	PE
i	0.600	0.167	0.000	0.000	0.000
ii	0.850	0.125	0.028	0.000	0.000
iii	0.400	0.500	0.000	0.125	0.000
iv	0.000	0.000	0.111	0.333	0.167
v	0.000	0.000	0.111	0.500	0.000
vi	0.022	0.000	0.489	0.089	0.044
vii	0.067	0.000	0.519	0.167	0.167
viii	0.000	0.000	0.444	0.000	0.000
ix	0.000	0.000	0.056	0.125	0.750

Verano

Grupos de especies	Grupos de arrastres A	B	C	D	E
i	0.000	0.000	0.083	0.708	0.333
ii	0.067	0.200	0.800	0.250	0.000
iii	0.833	0.444	0.000	0.000	0.000
iv	0.167	0.167	0.000	0.125	0.250

Analisis de correspondencias.

En las figuras 8 a 12 se muestra la localizacion de los arrastres y las especies en el plano formado por los dos primeros ejes del analisis de correspondencias de cada epoca del año y de todo el año. Se pueden distinguir ciertos grupos que se corresponden con algunos de los obtenidos en el analisis de cumulos. El primer eje separa a las especies y localidades de las partes interna y media de la externa de la plataforma continental en todos los casos. En la tabla 4 se dan los valores del coeficiente de correlacion de orden de Spearman de la primera coordenada de las localidades con la profundidad y la temperatura. Para la profundidad y la primera coordnada, la hipotesis de que el coeficiente de correlacion de orden es igual a 0 es rechazada con $p<0.05$ en todos los casos. Para la temperatura y la primera coordenada, la hipotesis se rechazo en el analisis de la primavera, el verano y global con $p<0.05$. Para las coordenadas de los otros ejes no se encontro correlacion con factores ambientales.

La separacion de las localidades y especies producida por el primer eje cambia de acuerdo a la epoca del año. En el analisis global la separacion que hacen los dos primeros ejes es muy similar a la de la primavera, pero com mas especies y localidades entre los dos grupos principales, estas son algunas de las que se encontraban situadas de manera similar en los resultados del verano, otoño e invierno. Sin embargo hay algunas especies y localidades que, encontrandose en posiciones intermedias en el primer eje, aparecen incorporadas a alguno de los grupos principales. Agrupando las coordenadas del primer eje de las localidades con menos de 100 m. de profundidad de cada estacion del año, se rechazo la hipotesis de que las cuatro muestras tienen la misma distribucion con $p<0.05$, con la prueba no parametrica que utiliza la distribucion de Kolmogorov. Para las localidades con mas de 100 m. de profundidad la hipotesis no se rechazo con $p=0.05$.

Al igual que en el analisis de cumulos, el analisis global no muestra una agrupacion de las localidades de acuerdo con las estaciones del año. Por otra parte, parece haber cierto agrupamiento debido a la zona de la plataforma donde se ubican; los arrastres realizados en la zona frente a las bahias de St. Andrew y Choctawatchee mostraron los valores mas altos del tercer eje.

La influencia de la temperatura es dificil de precisar ya que esta relacionada directamente con la profundidad. La mayor uniformidad de la fauna decapoda que se encontro en la plataforma externa, es probable que sea debida en parte a la estabilidad termica de la region a traves del año. La temperatura y la circulacion de las corrientes son

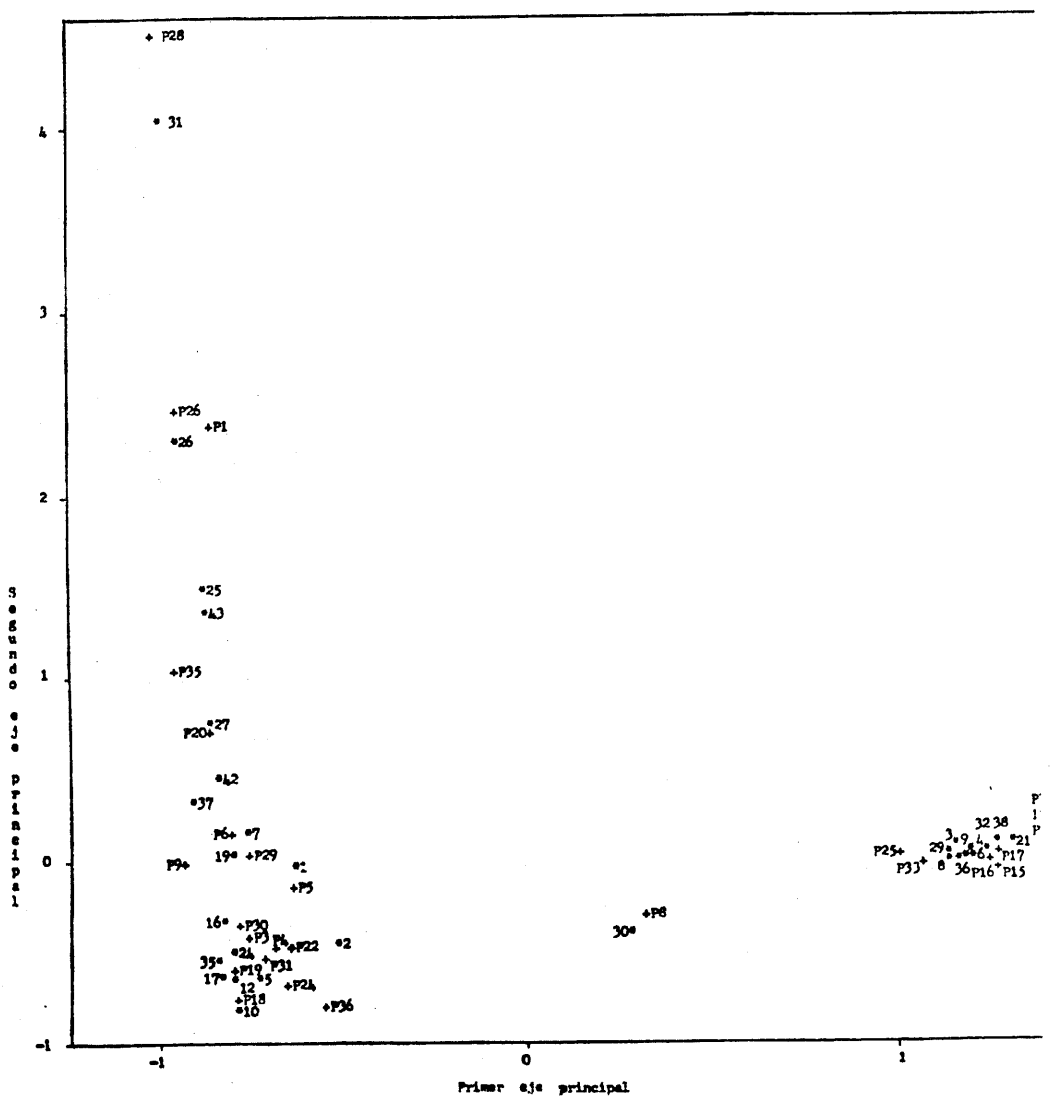

Segundo eje principal
Primer eje principal
4
3
2
1
0
-1
-1
0
1
+ P28
• 31
+P26
+P1
•26
•25
•43
+P35
P20+
•27
•42
•37
P6+
•7
19•
+P29
P9+
•
+P5
16•
+P30
+P3
+P22
•2
•24
35•
+P19
+P31
17•
•5
+P24
12
+P18
•10
+P36
+P8
30•
32 38
•21
P25+
29•
+P17
P33+
8
36
P16•
P15

fig. 8

Fig 61

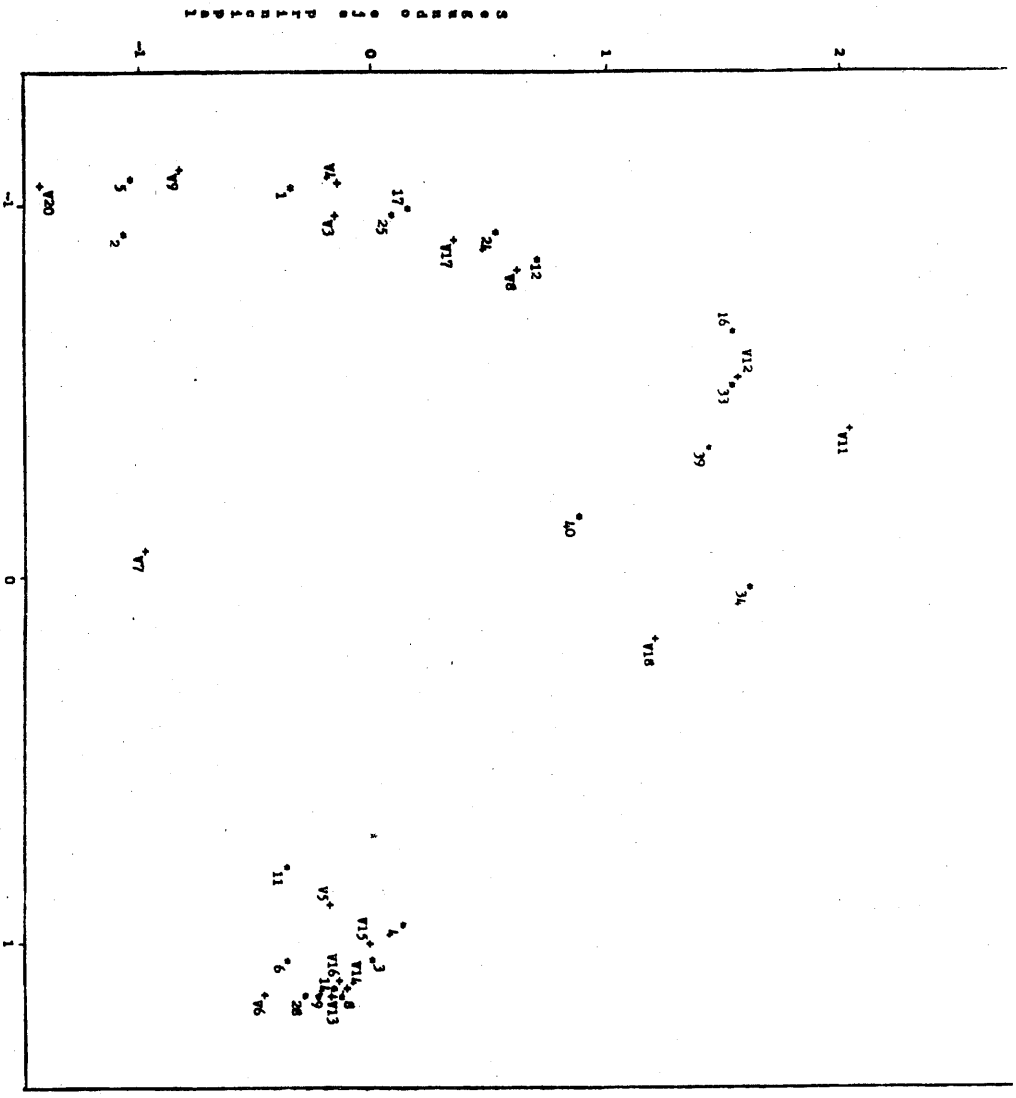

Segundo eje principal
-1
0
1
2
-1
0
1
V20
V9
5
2
1
V4
V3
17
25
24
V17
8
12
16
V12
33
V11
39
40
34
V18
V7
11
V5
V15
4
3
V14
8
V16
V13
7
28
6
V6

Fig. 9

Segundo eje principal
Primer eje principal
1
0
-1
12
5
6
13
3
11
111
16
22
15
112
2
110
18
1
14
16
19
17

fig. 44 //

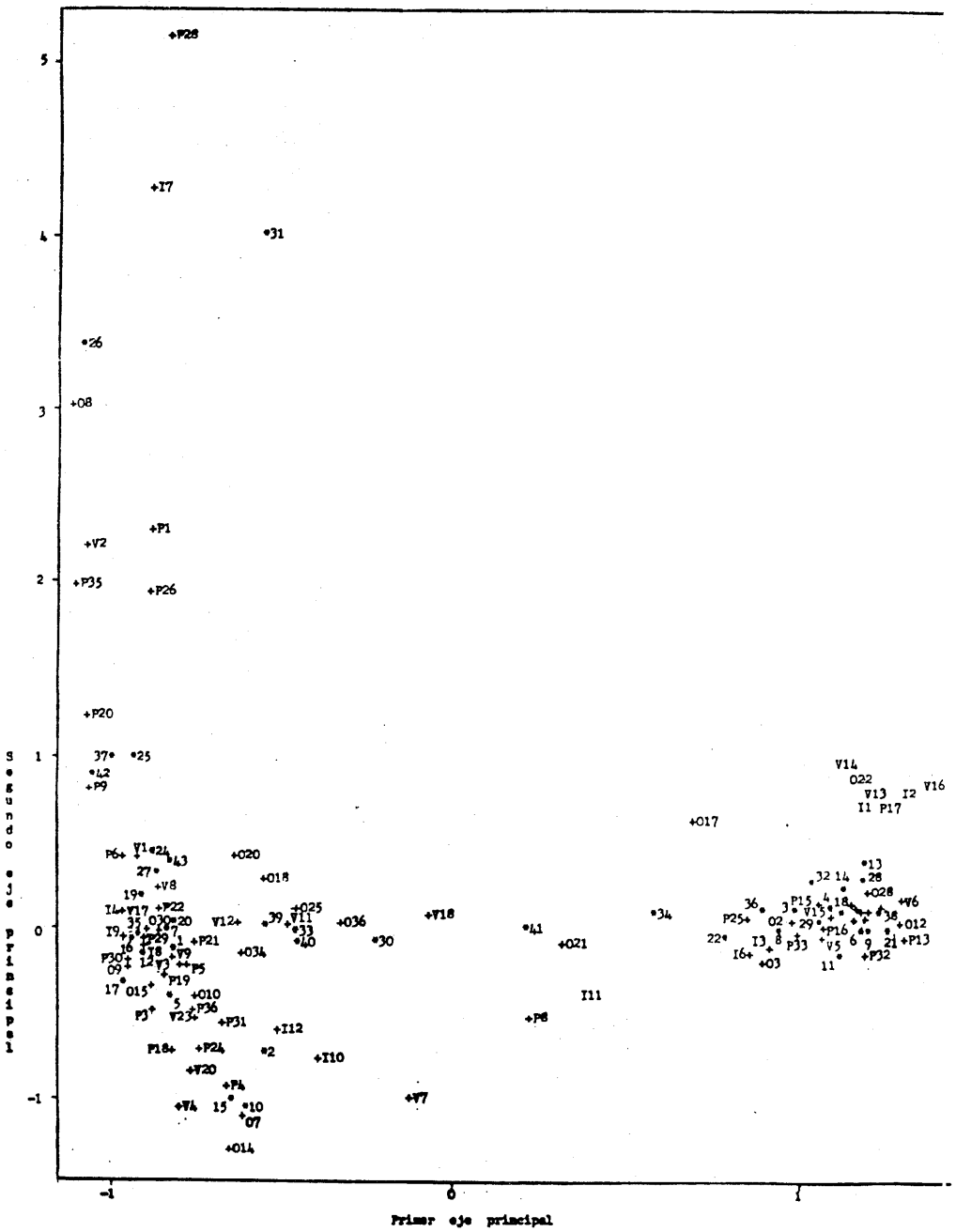


fig. 12

Tabla 4. Porcentaje de la varianza explicada por cada uno de los primeros cinco ejes y coeficiente de correlacion de orden de Spearman para la primera coordenada con la profundidad y temperatura.° N=numero de variables (especies).

Estacion	Porcentaje de la varianza explicada por el eje: 1	2	3	4	5	Coeficiente de correlacion de Spearman de la 1a. coordenada: Profundidad	Temperatura	N
Prim.	17.7	11.0	9.2	8.8	7.8	0.74224	0.70349	26
Verano	21.7	14.3	10.5	8.9	7.3	0.83095	0.71335	16
Otoño	18.7	16.1	15.1	12.7	8.7	0.63643	0.33260	18
Inv.(3)	40.0	23.7	11.3	9.5	7.6	0.66969	0.38180	10
Inv.(2)	27.0	18.9	16.5	12.7	9.5	0.80680	0.09550	11
Global	11.4	7.1	6.3	5.5	5.0	0.76403	0.55439	76

posiblemente los factores mas importantes en determinar la variacion estacional en la distribucion de la fauna decapoda. En el presente estudio el mayor cambio se registro en el numero de especies presentes en cada estacion del año. La diferencia registrada en la primera coordenada del analisis de correspondencias para las localidades, de acuerdo con la temporada del año, indica un translape gradual de la fauna de las dos componentes distinguidas, siendo mayor en el verano y en el otoño. De acuerdo con Wenner y Read (1982) este tipo de translape es una indicacion de cambios estacionales en la composicion de las comunidades. En general los cambios estacionales son dificiles de distinguir de los espaciales, particularmente cuando el estudio se hace solamente durante un año y cuando se utiliza un tipo de muestreo semicuantitativo, como es el de arrastres.

Indices de diversidad.

Los valores mas bajos del indice de diversidad se registraron en la zona cercana al Delta del Mississippi. La variacion de la diversidad estan en relacion directa con la riqueza de especies, independientemente de los cambios en la equitatividad.

Discusion de los metodos.

Anteriormente la clasificacion de las comunidades bentonicas se basaba en las especies dominantes que las caracterizaban y solamente se utilizaban unas pocas especies como atributos para clasificar. El uso de tecnicas de agupamiento jerarquico aglomerativo permite incorporar a un mayor numero de especies para ser usadas como atributos y con el uso de coeficientes como el de Camberra se puede evitar el sego hacia la dominancia. El analisis de cumulos proporciona un modelo de tipo de mosaico que resulta conveniente cuando la distribucion de los organismos presenta discontinuidades muy acentuadas. Por otra parte el analisis de correspondencias es un modelo continuo que se puede aplicar cuando existen gradientes que afectan a las comunidades, aunque la interpretacion no siempre es facil, ademas de que generalmente solo se pueden analizar las proyecciones del espacio en dos o tres dimensiones.

En el presente estudio ambos metodos se complementan por la gran diferencia que existe entre las dos componentes faunisticas distinguidas. Posiblemente la aplicacion del analisis de correspondencias a las partes interna y media por una parte y a la externa por otra, pueda facilitar la interpretacion de los resultados.

El gradiente de profundidad permite la utilizacion del analisis de correspondencias, sin embargo para otros

factores como la temperatura, el tipo de sedimento, las corrientes, la descarga de los rios, que varian tanto en el espacio como en el tiempo, la utilidad del analisis de correspondencias considerandolos no parece tan obvia, aunque posiblemente lo que se necesita es una mdida de estos factores que facilite la interpretacion.

Literatura citada.

BOESCH, D. F., 1973. Classification and community stucture of macrobenthos in the Hampton Roads area, Virginia. Mar. Biol., 21: 226-244.

FAGER, E. W. y J. A. McGOWAN, 1963. Zooplancton species groups in the north Pacific. Science, 140: 453-460.

MALMGREM, B., OVIATT C., GERBER R. y JEFFRIES H. P. 1978. Correspondence analysis: applications to biological oceanographic data. Est. Coast. Mar. Sci., 6: 429-437.

MARGALEF, R., 1958. Information theory in Ecology. Gen. Syst. 3: 36-71.

LANCE, G.N. y W.T. WILLIAMS, 1967. A general theory of classificatory sorting strategies. I. Hierarchical systems. Comput. J. 9: 373-380.

MUSICK, J.A. y McEACHRAM, 1972. Autumm and winter occurrence of decapod crustaceans in Chesapeake Bight, U.S.A. Crustaceana 22: 190-200.

PIELOU, E.C., 1966. The measurement of diversity in different types of biological collections. J. Theor. Biol. 13: 131-141.

REYES, L.A., LOPEZ y G. ESPINOZA,1978. Analisis/Cumulos. Un programa para el analisis de cumulos. Comunicaciones Tecnicas. Serie Amarilla: Desarrollo 6. Instituto de Investigaciones en Matematicas Aplicadas y Sistemas, U.N.A.M.

SOTO, L. A., 1972. Decapod shelf-fauna of the northeastern Gulf of Mexico, distribution and zoogeography. Unpubl. M. S. Thesis, The Florida State Univ. 122 p.

WENNER, E. L. Y READ, T. H., 1982. Seasonal composition and abundance of decapod crustacean assamblages from the South Atlantic Bight, USA. Bull. Mar. Sci., 32 (1): 181-208.

ESTUDIO COMPARATIVO DEL POTENCIAL CIENTIFICO Y TECNOLOGICO DE MEXICO+ Y HUNGRIA

Jaime Jimenez*
Peter Hunya**
Magdalena Bayona*
Arpad Halasz***

*Instituto de Investigaciones en Matematicas Aplicadas y en Sistemas.
Universidad Nacional Autonoma de Mexico.

**Laboratorio de Cibernetica Kalmar.
Universidad Jozsef Attila.
6720 Szeged.
Arpad ter 2.
Hungria.

***Instituto de Administracion Cientifica en Informatica
Ministerio de Cultura e Information.
1111 Budapest. Egri Jozsef u. 1-9. "E"
Hungria.

Una version anterior de este trabajo fue presentada en el XI Congreso Internacional de Sociologia de la Asociación Internacional de Sociologia. bajo el titulo: "The S & T Potential of Mexico and Hungary". Nueva Delhi. India. Agosto 18-22. 1986.

+Este proyecto es patrocinado en Mexico por el Gobierno Federal a través del Consejo Nacional de Ciencia y Tecnologia (CONACYT).

R E S U M -

1. Introduccion

México y Hungria estan participando simultáneamente en el cuarto ciclo del Estudio Comparativo Internacional Sobre la Organizacion y Eficacia de las Unidades de Investigación (ICSOPRU), que es un proyecto coordinado por la UNESCO a nivel internacional. Este proyecto se basa en la realizacion de encuestas nacionales sobre un universo de unidades de investigación, definido por cada pais participante. Los cuestionarios aplicados son iguales para todos los paises. Los responsables del proyecto en Mexico y Hungria decidimos que resultaria muy interesante comparar el potencial cientifico y tecnologico de ambos paises, aprovechando que se cuenta con datos recientes de poblaciones cientificas comparables. A continuacion se enuncian algunas de las caracteristicas mas importantes de ambos paises:

Hungria se encuentra ubicada en el centro de Europa, con una densidad de poblacion aproximada de 108 personas por Km2 y una poblacion total de 11 millones de habitantes. Mexico situado en el sur de Norteamerica, cuenta con una densidad de poblacion menor de 41 personas por Km2, y una poblacion total de aproximadamente 80 millones de habitantes, distribuida desigualmente a lo largo del territorio nacional.

La alta tasa de natalidad que ha tenido Mexico, en el pasado reciente (3.5% anual en 1970), ha bajado aproximadamente a 2.1% para 1986. Su poblacion es muy joven, ya que un 60% del total son menores de 21 años. Esto contrasta con la distribucion de edades y la tasa de nacimientos en Hungria, tipicas de un pais centroeuropeo. Ambos paises muestran una alta concentración de población en la capital, aunque el problema es menor en Hungria con 2 millones de habitantes en Budapest y sus alrededores, contrastado con unos 18 millones de habitantes en el area metropolitana de la ciudad de México. En términos de la productividad nacional, en 1985 el PIB per capita mexicano fué de $ 2276.00 - Dlls., y en Hungria fue de $ 1512.00 Dlls.

La ubicación geográfica de Hungria permite frecuentes intercambios cientificos y tecnológicos con paises vecinos de Europa oriental y occidental. En Mexico, la mayor parte de los intercambios de ciencia y tecnología (C y T) tienen lugar con Estados Unidos y en menor proporción con paises europeos. Paradójicamente, aunque México es un pais latinoamericano, tiene menor intercambio de ciencia y tecnologia con otros paises latinoamericanos que con Estados Unidos y Europa.

2. Metodolcgia

El proyecto ICSOPRU centra su atencion en el desempeño de las unidades de investigacion, definidas como un grupo que cuenta cuando menos con tres miembros, incluido el jefe, formado por los cientificos, ingenieros y técnicos. Dicho equipo debe estar trabajando en cuando menos un proyecto que tenga una duracion minima de un año.

Se obtuvo informacion a tres niveles jerárquicos: del director de la institucion de la que forma parte la unidad, del jefe de la unidad, y de los miembros de la unidad. Este trabajo analiza las respuestas de los directores de institucion y de los jefes de unidad.

Marcos muestrales

México

El universo del estudio fue definido como:

> todas aquellas unidades que pertenecen a instituciones cuyo principal objetivo es la investigacion cientifica o -- tecnológica. (Jimenez, 1986).

Basados en el inventario de Ciencia y Tecnologia hecho por el Consejo Nacional de Ciencia y Tecnologia (CONACYT .1985), fueron identificadas un total de 247 instituciones cuya actividad principal es la investigación cientifica o tecnológica. Se segmentó el universo de instituciones de acuerdo al sector de pertenencia en los estratos siguientes:

- Instituciones del Gobierno Federal (52)
- Instituciones académicas públicas (171)
- Instituciones académicas privadas (15)
- Otras instituciones (empresas privadas, asociaciones no lucrativas, organizaciones internacionales, - etc.) (9)

Por otro lado, debido a las diferencias importantes (organización, tamaño, nivel académico, presupuesto) entre las instituciones del área metropolitana de la ciudad de México, ("centro") y las de provincia, ("periferia") el universo también se dividió en instituciones del D. F. e instituciones de provincia. La Tabla 1 muestra la distribución del universo de instituciones por estrato y ubicación geográfica.

Tomando en cuenta que el número de instituciones definidas para este estudio es pequeño (247), se incluyeron a todas, salvo en el caso de las "académicas públicas-periferia" Para este estrato seleccionamos al azar un numero representativo (72), que corresponde a una razon de muestreo de 0.42.

Como las unidades de investigacion presentan mayores diferencias "entre" instituciones que "dentro" de las mismas, con objeto de obtener la mayor variedad posible, cada institucion de la muestra fué representada por al menos una unidad de investigacion. Finalmente, se encuestaron 114 unidades en la zona metropolitana de la ciudad de Mexico, y 107 fuera de ella. El numero total de unidades que respondieron la encuesta fue de 221. La Tabla 2, muestra la distribucion de las unidades encuestadas por estrato.

Hungria

En Hungria el potencial de ciencia y tecnologia esta basicamente localizado en dos estratos, instituciones pertenecientes al sector educativo (universidades, institutos tecnologicos, instituciones de educacion superior, etc..., e instituciones pertenecientes a la Academia de Ciencias.

Hungria participo hace algunos años en el primer ciclo de ISOPRU, dandole preferencia a los centros de investigacion pertenecientes a la Academia de Ciencias. En el cuarto ciclo se selecciono el universo de estudio de tal manera que un mayor numero de instituciones del sector educativo apareciera en la muestra.

En la selección de unidades del sector educativo, fué utilizado un procedimiento de muestreo en dos etapas. Primero se obtuvo una muestra de 190 instituciones de un universo de 927 instituciones de investigación del sector. Después se seleccionaron al azar 266 unidades de un total de 413 que componen la muestra institucional.

Similarmente, las unidades que no pertenecen al sector educativo fueron muestreadas con un procedimiento en dos etapas. Primero, del total de instituciones de investigación no pertenecientes al sector educativo (225), fueron seleccionadas al azar 42. A continuación se obtuvo una muestra aleatoria de 75 unidades del total de 517 de la muestra institucional. Finalmente debido a que algunas unidades no contaban con el mínimo de integrantes (tres) y otras se rehusaron a participar, la muestra se redujo a 222 unidades, 155 del sector educativo y 67 del sector no educativo. Las Tablas 3 y 4 sintetizan el procedimiento muestral empleado.

Comparabilidad de los marcos muestrales

Todas las instituciones cuya labor prioritaria es la investigación científica o el desarrollo tecnológico, fueron incluidas en el universo de estudio mexicano. De manera similar, las instituciones de C y T húngaras están representadas adecuadamente, ya que todo tipo de unidad de investigación está incluida en el universo de estudio.

En términos de la distribución por tipo de unidad, ambas muestras manejan cantidades comparables. La Tabla 1 destaca que ambas tienen la misma proporción de unidades educativas y no educativas en la muestra final. Es verdad que el diseño de la muestra húngara favorece la selección de unidades en el sector educativo. Sin embargo, después del primer paso del procedimiento muestral, el equipo húngaro seleccionó una de cada 2.7 unidades del sector educativo y una de cada 7.7 unidades del sector no educativo. Esto no está fuera de línea con el tamaño de la muestra final mexicana que seleccionó una de cada 2.2 unidades del sector educativo, y una de cada 5.6 unidades de otros sectores. Dado que el propósito de este estudio es solamente señalar grandes diferencias del potencial de ciencia y tecnología entre ambos países, no fue necesario ponderar la información obtenida por ambas encuestas.

3. Análisis

Es evidente que la ciencia en Hungría está en una etapa más avanzada que en México. La infraestructura científica y tecnológica en Hungría es mucho mayor que en México tanto en términos relativos como absolutos. Comparando las Tablas 1 y 3, se puede observar que Hungría tiene un mayor número de instituciones científicas, por un factor de cuatro, que México. Desde luego, Hungría tiene más personal científico que nuestro país (ver Figura 1). La investigación científica, entendida como un cuerpo organizado de individuos dedicados tiempo completo a la investigación, es un fenómeno mucho más antiguo en Hungría que en México. El 54% de las instituciones húngaras que fueron encuestadas se fundaron antes de 1951, contrastando con solamente el 21% de las instituciones mexicanas, como lo ilustra la Figura 2. En términos de personal científico altamente calificado, Hungría tiene más científicos doctorados que México, como se aprecia en la Figura 3, referida a las instituciones, y en la Figura 5, referida a las unidades de investigación. Esta afirmación puede ser un tanto matizada por la naturaleza de los grados doctorales que otorga usualmente Hungría.

Las unidades de investigación húngaras llevan a cabo una cantidad substancial de investigación relacionada con los problemas en el sector productivo, mientras que las unidades mexicanas no tienen fuertes ligas con tan importante sector. Esta afirmacion se apoya en los resultados mostrados en la Figura 4 (instituciones), y en la Figura 6 (unidades). En consecuencia, las unidades húngaras consiguen más apoyo económico de la industria que las mexicanas, como se muestra en la Figura 7. De hecho, las unidades mexicanas son financiadas fundamentalmente con fondos del Gobierno Federal (Figura 8), y otras fuentes de financiamiento juegan un papel menor en la configuracion de su presupuesto. La ciencia mexicana, parece estar mas orientada hacia las ciencias naturales, y hacia las ciencias sociales y las humanidades, como se puede apreciar observando los porcentajes de la Figura 4 para las instituciones, y de la Figura 6, para las unidades. La etapa actual del desarrollo en Mexico parece estar reflejada por la tendencia de la ciencia de cumplir objetivos referentes a las ciencias sociales (ver Figura 5).

En terminos de productividad impresa, las unidades mexicanas publicaron más libros que las húngaras en el mismo periodo de tiempo (36 meses), segun se muestra en la Figura 10. En contraste, las unidades húngaras publicaron más artículos de investigacion dentro del país que las mexicanas, de acuerdo a la Figura 11. El número de publicaciones de las unidades húngaras en el extranjero es mas grande que el de las unidades mexicanas, como se puede constatar en la Figura 12. Esta caracteristica puede ser interpretada como una medida de la orientación internacional de la ciencia y la tecnología húngaras. De la misma manera, el alto porcentaje de unidades mexicanas que no publican en el extranjero (81%) puede ser tomado como una medida de la orientacion menos internacional de la ciencia y tecnología mexicanas. Aunque un mayor numero de investigadores mexicanos no publican libros o artículos, un segmento de los mismos mantiene una produccion estable, tanto en Mexico como en el extranjero. En efecto, cierto numero de unidades mexicanas compite favorablemente con las unidades húngaras en el rango de alta productividad, como se aprecia en las Figuras 10, 11 y 12. Un mayor numero de unidades mexicanas no recibe solicitudes de consultas cientificas o servicios técnicos. Como en el caso de las publicaciones, un segmento de unidades mexicanas sí recibe solicitudes regularmente y compite con las unidades húngaras en este respecto.

Es interesante hacer notar que en términos de recursos para el trabajo futuro ambos paises muestran un alto nivel de necesidades, en particular respecto a un mayor financiamiento. Hungria refleja mayor necesidad de equipo de laboratorio especializado, lo cual puede ser interpretado como la existencia de un aparato cientifico más sofisticado que el de México. Las unidades mexicanas mostraron una gran necesidad de reforzar la accesibilidad a la literatura cientifica nacional. Esto se puede interpretar como un aspecto de infraes-

tructura basica que aún es raquítica en México. La Tabla 6 (p. 15) sintetiza las diferencias más importantes entre el potencial científico y tecnológico de México y Hungría, de acuerdo a los análisis presentados en este estudio. Por brevedad, no se muestran todas las gráficas en que se basa este análisis (ver Jiménez, Hunya, Bayona, Halász, 1986).

Es necesario llevar a cabo una investigación más a fondo para entender las diferencias fundamentales entre los sistemas de ciencia y tecnología de Hungría y de México. Sin embargo, este estudio preliminar ya hace patentes algunas características interesantes de ambos sistemas. La importancia de este estudio radica en el hecho de se cuenta con datos recientemente registrados y con un alto nivel de confiabilidad.

Agradecimientos

Los autores desean agradecer a Martha W. Rees la revisión del manuscrito final, a Jorge Domínguez la elaboración de los programas de computadora, y a Francisco J. Jiménez la producción de las figuras.

Referencias

CONACYT. Directorio Nacional de Instituciones y Universidades que realizan Investigación y Desarrollo Experimental. Consejo Nacional de Ciencia y Tecnología. Centro Cultural Universitario. México, 1985.

Jiménez, J. Proyecto ICSOPRU-CONACYT: Resultados Preliminares. Reporte para el Comité Asesor. IIMAS, UNAM. México, 1986.

Jiménez, J., Hunya, P., Bayona, M., Halász, A. "The S & T Potential of Mexico and Hungary". Comunicaciones Técnicas. Serie naranja: investigaciones. No. 433. IIMAS, UNAM. México, 1986.

TABLAS Y FIGURAS

SECTOR	CENTRO	PERIFERIA	TOTAL	%
GOBIERNO FEDERAL	31	21	52	21%
ACADEMICAS PUBLICAS	54	117	171	69%
ACADEMICAS PRIVADAS	4	11	15	6%
OTRAS	3	6	9	4%
TOTAL	92	155	247	100%
%	37%	63%	100%	

Tabla 1. Distribución del universo de instituciones mexicanas por sector y localización geográfica.

SECTOR	CENTRO	PERIFERIA	TOTAL	%
GOBIERNO FEDERAL	37	23	60	27%
ACADEMICAS PUBLICAS	65	72	137	62%
ACADEMICAS PRIVADAS	9	9	18	8%
OTRAS	3	3	6	3%
TOTAL	114	107	221	
%	52%	48%		100%

Tabla 2. Número de unidades encuestadas en México por estrato.

SECTOR	UNIVERSO DE INSTITUCIONES	%	INSTITUCIONES EN LA MUESTRA	%
EDUCATIVO	927	80%	190	82%
OTRAS	225	20%	42	18%
TOTAL	1152	100%	232	100%

Tabla 3. Primera etapa del procedimiento muestral del potencial científico y tecnológico de Hungría.

SECTOR	UNIVERSO DE UNIDADES	%	UNIDADES EN LA MUESTRA	%	UNIDADES ENCUESTADAS	%
EDUCATIVO	413	44%	266	78%	155	70%
OTRAS	517	56%	75	22%	67	30%
TOTAL	930	100%	341		222	100%

Tabla 4. Segunda etapa del procedimiento de la muestra del potencial húngaro de C y T.

	NUMERO DE UNIDADES		
SECTOR	MEXICO	HUNGRIA	%
EDUCATIVO	155	155	70%
OTRAS	66	67	30%
TOTAL	221	222	100%

Tabla 5. Número de unidades de investigación muestreadss por sector en México y Hungría.

CARACTERISTICAS DE LAS INSTITUCIONES

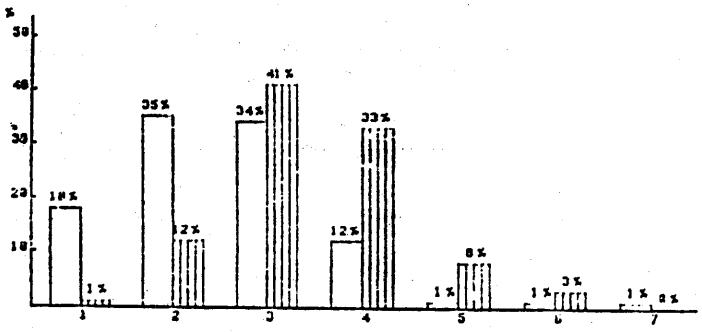


Figura 1. Personal de investigación por institución (científicos, ingenieros y técnicos).

Clave:

1: 3 a 12
2: 13 a 50
3: 51 a 200
4: 201 a 600
5: 601 a 1000
6: 1001 a 3000
7: Más de 3000

□ México

▨ Hungría

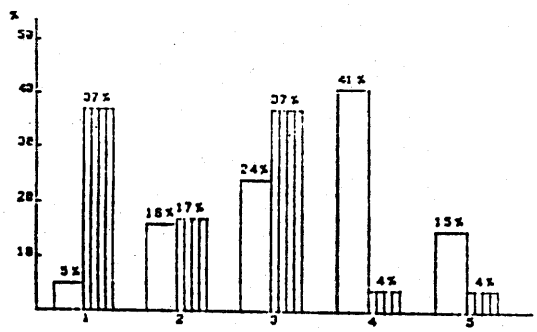


Figura 2. Período en que fué creada la institución.

Clave:

1: En 1930 o antes
2: 1931 - 1950
3: 1951 - 1970
4: 1971 - 1980
5: Después de 1980

□ México

▨ Hungría

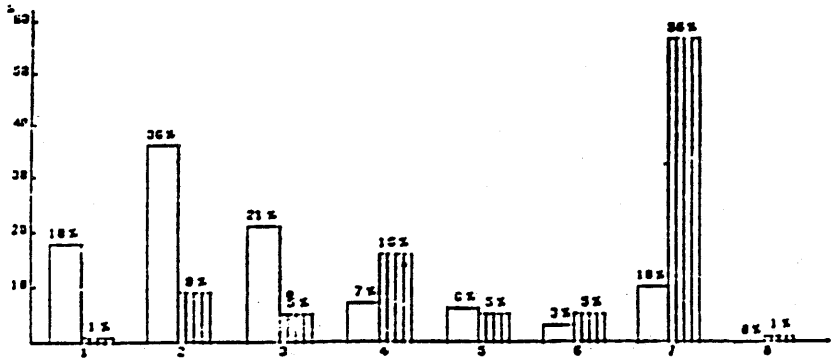


Figura 3. Proporción de investigadores con doctorado respecto a los científicos e ingenieros, - por institución.

Clave:

1:	Cero doctores	5:	31 - 40%
2:	1 - 10%	6:	41 - 50%
3:	11 - 20%	7:	Mayor que 50%
4:	21 - 30%	8:	Sin respuesta

□ México

▥ Hungría

ORIENTACION DE LA INVESTIGACION
DENTRO DE LAS INSTITUCIONES

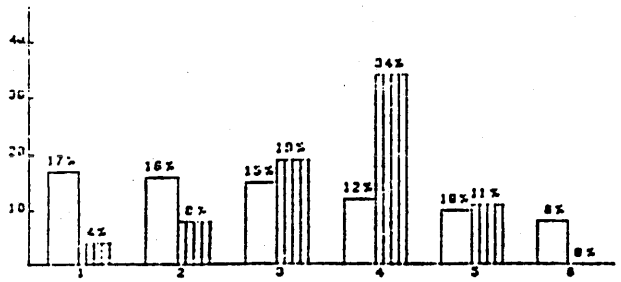


Figura 4. Importancia de las actividades de investigación y desarrollo que otorga la institución a los siguientes objetivos.

Clave:

1: Desarrollo social y servicios socio-económicos.
2: Adelanto general del saber.
3: Desarrollo de los servicios educativos.
4: Promoción del desarrollo industrial.
5: Desarrollo de los servicios de salud.
6: Exploración y estimación de los recursos de la tierra, los mares, la atmósfera y el espacio.

CARACTERISTICAS DE LAS UNIDADES DE INVESTIGACION

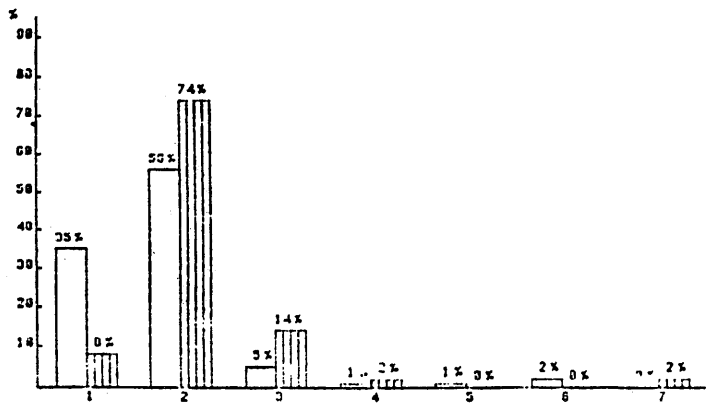


Figura 5. Número de científicos con doctorado en la unidad.

Clave:

1: Cero doctores
2: 1 - 5
3: 6 - 10
4: 11 - 15
5: 16 - 20
6: 21 o más
7: Sin respuesta

□ México
▥ Hungría

FUENTES DE INFLUENCIA

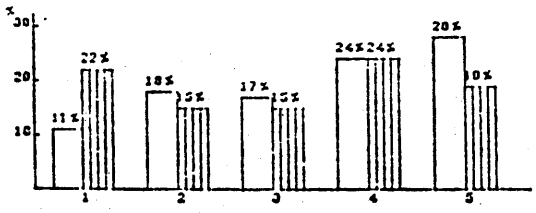


Figura 6. Importancia concedida a las necesidades del sector productivo en la elección de temas de investigación.

Clave:

1: La más importante
2: La segunda en importancia
3: La tercera en importancia
4: La cuarta en importancia
5: La quinta en importancia

□ México
▥ Hungría

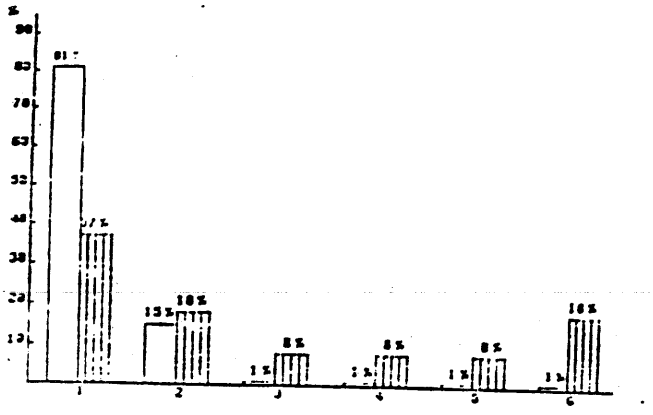


Figura 7. Porcentaje de fondos provenientes del sector productivo utilizados por la unidad en los últimos 36 meses.

Clave:

1: Cero por ciento
2: 1 - 20%
3: 21 - 40%
4: 41 - 60%
5: 61 - 80%
6: 81 - 100%

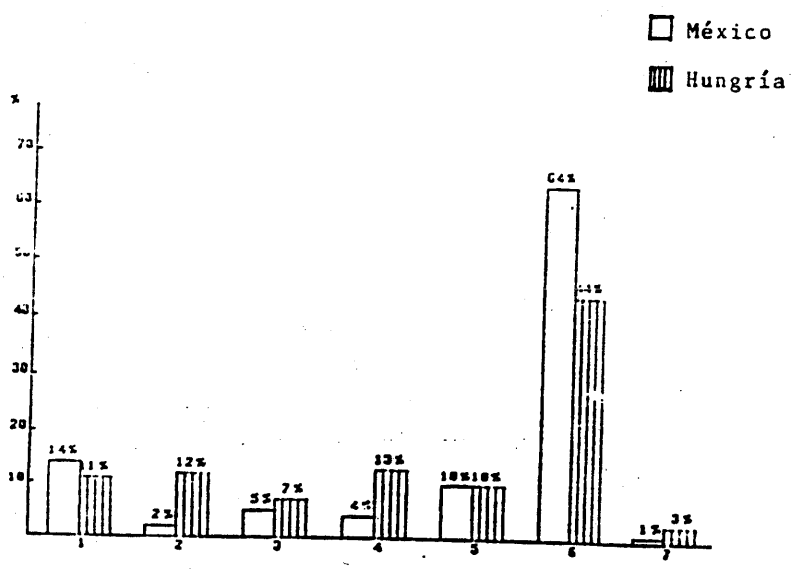


Figura 8. Porcentaje de fondos provenientes del gobierno utilizados por la unidad en los últimos 36 meses.

Clave:

1: Cero por ciento
2: 1 - 20%
3: 21 - 40%
4: 41 - 60%
5: 61 - 80%
6: 81 - 100%
7: Sin respuesta

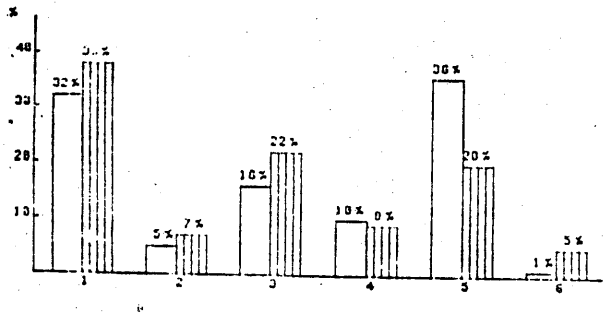


Figura 9. Campos de C y T donde la unidad debería concentrar sus esfuerzos en los próximos 4 a 6 años.

Clave:

1: Ciencias naturales
2: Ciencias y tecnologías agropecuarias
3: Ciencias y tecnologías de la ingeniería
4: Ciencias y tecnologías de la salud
5: Ciencias sociales y humanidades
6: Sin respuesta

□ México

▥ Hungría

PUBLICACIONES

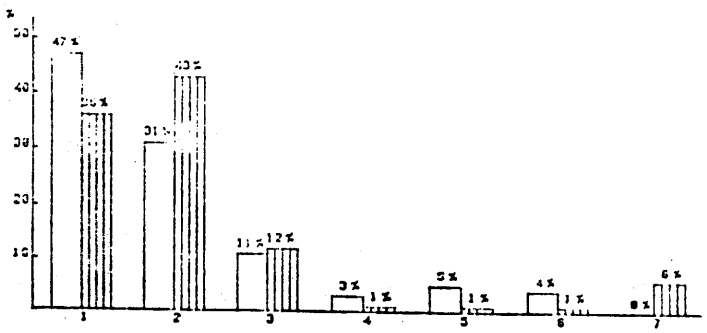


Figura 10. Número de libros publicados por cada científico e ingeniero en la unidad en los últimos 36 meses.

Clave:

1: Cero libros
2: 0 - 0.5 libro
3: 0.5 - 1.0 libro
4: 1.0 - 1.5 libros
5: 1.5 - 2.0 libros
6: Más de dos libros por científico e ingeniero
7: Sin respuesta

□ México

▥ Hungría

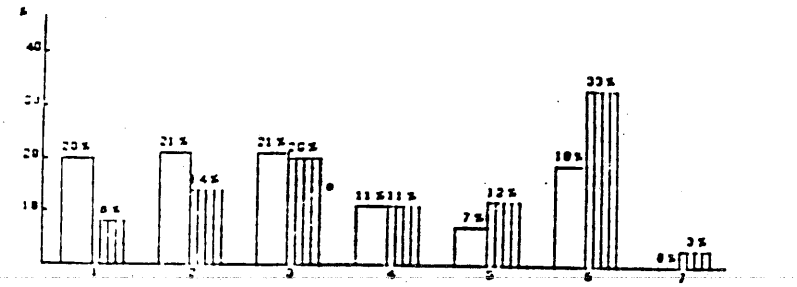


Figura 11. Número de artículos científicos o técnicos publicados en el país por cada científico e ingeniero en los últimos 36 meses.

Clave:

1: Cero artículos
2: 0 - 0.5 artículo
3: 0.5 - 1.0 artículo
4: 1.0 - 1.5 artículos
5: 1.5 - 2.0 artículos
6: Más de dos artículos por científico e ingeniero
7: Sin respuesta

□ México
▥ Hungría

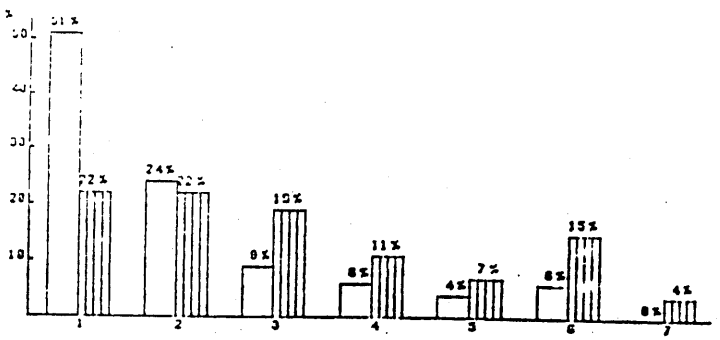


Figura 12. Número de artículos científicos o técnicos publicados en el extranjero por cada científico e ingeniero en los últimos 36 meses.

Clave:

1: Cero artículos
2: 0 - 0.5 artículo
3: 0.5 - 1.0 artículo
4: 1.0 - 1.5 artículos
5: 1.5 - 2.0 artículos
6: Más de dos artículos por científico e ingeniero
7: Sin respuesta

□ México
▥ Hungría

MEXICO	HUNGRIA
Las instituciones y unidades son más jóvenes	Las instituciones y unidades son más antiguas
Infraestructura de C y T - pequeña	Infraestructura de C y T grande
Menor número de personal - científico calificado	Mayor número de personal científico calificado
Investigación orientada hacia el sector social	Investigación orientada hacia el sector productivo
Mayor producción de libros	Menor producción de libros
Menor número de trabajos - científicos publicados en el país y en el extranjero	Mayor número de trabajos científicos publicados en el país y en el extranjero
Mayor necesidad de personal calificado	Mayor necesidad de personal - calificado
Necesidad de mayor financiamiento	Necesidad de menor financiamiento
Menor necesidad de equipo - de laboratorio especializado	Mayor necesidad de equipo de laboratorio especializado

Tabla 6. Diferencias principales del potencial científico y tecnológico de México y Hungría.

MODELADO DE TRAFICO TELEFONICO

Dra. Teresa López A.
*Gerencia de Estudios Económicos
TELEFONOS DE MEXICO, S.A. DE C.V.

Para el proceso de planeación de Teléfonos de México es importante conocer los ingresos que se van a generar y estos se obtienen principalmente en base al tráfico telefónico. Aproximadamente el 80% de los ingresos son generados por tráfico de larga distancia tanto nacional como internacional.

De aquí la relevancia de pronosticar el tráfico telefónico de larga distancia y los ingresos generados por éste.

El tráfico de larga distancia se divide en 3 tipos:

i) Larga distancia nacional, que es aquel que se genera dentro del territorio nacional y no se involucran administraciones extranjeras.

ii) Larga distancia internacional cobrado en México. Es tráfico que puede generarse en el extranjero o en nuestro país con destino fuera de éste, pero que se cobra en México.

iii) Larga distancia internacional cobrado en el extranjero. Este tráfico puede tener su origen en México o fuera de México pero es cobrado en el extranjero.

El objetivo de este trabajo es presentar el modelo que se diseñó para pronosticar ingresos generados por el tráfico de larga distancia internacional cobrado en México (LDICM).

El tráfico se puede medir a través del número de conferencias o bien del número de minutos.

Tradicionalmente en la empresa se modelaba el número de conferencias; se estimaba por gente experimentada el ingreso por conferencia y de aquí se obtenían ingresos.

Este enfoque falló a partir de 1983 ya que a raíz de la crisis las conferencias duraban menos tiempo que antes y esto repercutió en los ingresos, pues la facturación se realiza por tiempo.

De aquí que para pronosticar los ingresos se considerarán los minutos como variable de interés. Para este fin se utilizan los siguientes dos modelos:

1) Modelo simultáneo de minutos y conferencias

2) Modelo de ingresos por minuto

Sólo detallaremos en este trabajo el modelo simultáneo. Ambos son modelos lineales.

2. Modelo Simultáneo de Minutos y Conferencias

Cabe señalar que el propósito de este modelo es el de pronóstico, aunque también nos interesa posteriormente hacer algún análisis sobre elasticidades.

Este modelo incluye una ecuación para conferencias, ya que esta sigue siendo una variable de interés para la empresa.

Ecuación de Conferencias

Las variables que intervienen en esta ecuación son:

CONF = (IMP, LINEAS, INCO, V.M.)

donde:

CONF = Número de conferencias generadas por tráfico internacional cobradas en México.

IMP = Importaciones totales de mercancías

LINEAS = Números de líneas telefónicas instaladas acumuladas al mes

INCO = Indice de precios Divisia generado con base a ingresos por conferencia y considerando los 4 tipos de servicio

T-T	Teléfono a teléfono	servicio manual
P-P	Persona a persona	
L-T	LADA teléfono a teléfono	servicio automático
L-P	LADA persona a persona	

V.M. = Variables mudas estacionales y variable para efectos del terremoto de 1985

Las importaciones son incluidas en la ecuación ya que el tráfico internacional depende en gran parte del intercambio comercial que exista entre México y otros países. El índice es considerado como la variable precio y de aquí se puede analizar la elasticidad del precio del servicio. Es importante incluir las líneas en el modelo ya que es una variable sobre la cual la empresa tiene control. El tráfico telefónico es estacional, y para tomar en cuenta esto se incluyen las variables mudas.

Ecuación de Minutos

Las variables que se incluyen en esta ecuación son:

MTOT = (CONF, PIB, INMI, V.M.)

donde:

MTOT = Número de minutos facturados al mes en llamadas de LDI cobradas en México

CONF = Número de conferencias LDICM

PIB = Producto Interno Bruto Real

INMI = Indice de precios Divisia generado con base en ingresos por minuto y considerando los 4 tipos de servicio mencionados en la ecuación 1.

V.M. = Variables mudas estacionales y variable para efectos del terremoto de 1985.

Las conferencias se incluyen en la ecuación de minutos ya que para que se generen minutos es necesario que primero se realice la conferencia. Esta variable es considerada como variable endógena. El producto interno bruto es la variable ingreso del modelo y el índice INMI la variable precio, además para tomar en cuenta el patrón estacional se consideran variables mudas.

Especificación del Modelo

El modelo queda especificado de la siguiente forma:

$$1)\ CONF_t = a_o + a_1\ IMP_t + a_2\ INCO_{t-1} + a_3\ LINEAS_t + {}^{s}V.M.{}^{s} + e_{1t}$$

$$2)\ MTOT_t = b_o + b_1\ CONF_t + b_2\ PIB_{t-1} + b_3\ INMI_{t-1} + {}^{s}V.M.{}^{s} + e_{2t}$$

donde las variables endógenas son CONF y MTOT, y las variables exógenas externas son PIB, IMP y ${}^{s}V.M.{}^{s}$; y las variables exógenas de política son los índices INMI, INCO y las LINEAS.

Las formulación se realizó pensando que existe correlación contemporánea entre MTOT y CONF, es decir que COV $(e_{1t}, e_{2t}) \neq 0$. Además, la ecuación (2) tiene una variable endógena en el lado derecho de la igualdad. Tomando en cuenta estos dos aspectos se utilizó como método de estimación mínimos cuadrados trietápicos.

Antes de realizar la estimación se verificó que el modelo fuera identificable. De hecho es un modelo sobreidentificado ya que cada ecuación tiene 4 parámetros de interés y 6 variables exógenas, esto es sin incluir las variables mudas.

Estimación

La estimación del modelo se realizó utilizando datos mensuales de agosto de 1976 a mayo de 1986.

Ecuación (1)

$$CONF_t = \underset{(-3.31)}{-325.7} + \underset{(10.07)}{0.27442}\ IMP_t - \underset{(-5.3)}{8641.7}\ INCO_{t-1} + \underset{(39.3)}{.00073793}\ LINEAS_t + (etc)$$

$R^2 = .951$ D.W. = 0.80

Ecuación (2)

$$MTOT_t = \underset{(-1.78)}{-2823900} + \underset{(6.54)}{3323.7}\ CONF_t + \underset{(5.58)}{86.901}\ PIB_{t-1} \underset{(-5.91)}{-68958000}\ INMI_{t-1} + (etc.)$$

$R^2 = .949$ D.W. = 0.28

Se observa que en ambas ecuaciones los signos son los correctos; los índices de precios tienen signos negativos, mientras que PIB, IMP y LINEAS tienen signo positivo. Para todas las variables de interés el estadístico 't' resulta signficativo. Los coeficientes de determinación están alrededor de 0.95. El único problema que presenta el modelo es el de autocorrelación.

Pronósticos

Para realizar los pronósticos es necesario conocer las premisas de las variables exógenas del modelo. Los supuestos a futuro sobre PIB e IMP son obtenidos del escenario económico de Telmex. Las líneas son propuestas por otra área de la empresa y los pronósticos sobre los índices se generan usando un modelo que tiene como variable exógena la paridad controlada, y esta a su vez es obtenida del escenario.

Con base en estas premisas y usando las ecuaciones estimadas se generaron los pronósticos de minutos y conferencias. Estos pronósticos resultaron satisfactorios, ya que son congruentes con las expectativas de gente experimentada en esta área, además se analizaron algunos otros indicadores (por ejemplo minutos x conferencia) y también presentan resultados aceptables.

CONCLUSIONES

La generación de pronósticos que van a ser utilizados en la toma de decisiones es un proceso largo que incluye la especificación de varios modelos; hasta obtener resultados que, por una parte tengan un sustento teórico aceptable y que por otro lado sean satisfactorios (tengan sentido para la gente conocedora) para el usuario.

En este trabajo se introdujo la duración de las llamadas para la obtención de los pronósticos de ingresos, a través de un modelo simultáneo de minutos y conferencias. En la estimación del modelo se toma en consideración que exista correlación contemporánea entre minutos y conferencias.

Aquí se presentó un modelo preliminar, que no se puede considerar como la versión final, pero que ya genera pronósticos aceptables que son utilizados en la práctica.

ESTADISTICA Y AGRONOMIA

Ing. Agr. y Dr. Ignacio Méndez R.

CONTENIDO:

ESTADISTICA Y AGRONOMIA*

Ing. Agr. y Dr. Ignacio Méndez R.

I Introducción

En este breve escrito, se presenta un panorama de las relaciones entre la agronomía y la estadística. Se inicia con algunos comentarios de tipo histórico, donde resaltan las contribuciones de Fisher. Sin embargo, se presentan dos críticas a sus ideas. La primera por la falta de homogeneidad en la produtividad de las parcelas agrícolas dentro de un bloque y la segunda porque los bloques mismos no tienen repeticiones independientes. Finalmente se comentan otros desarrollos de la estadística y la agronomía de manera conjunta, en estimulación recíproca.

En un tema tan amplio como éste, es imposible señalar todos los innumerables trabajos que aplican o desarrollan técnicas estadísticas usadas en el avance de la agronomía.

II Comentarios sobre aspectos Históricos

A medida que la investigación científica fue estudiando objetos que presentan más y más variabilidad, por ejemplo el pasaje de la óptica, cinemática e hidráulica a la biología y la agronomía; se hizo muy urgente el problema de diseñar experimentos que controlen esta variabilidad de manera explócita. Sin embargo a menudo, los problemas urgentes de una disciplina no son reconocidos ampliamente por los practicantes de la misma. Así sucedió con la agronomía, en relación al problema señalado, el del control de la enorme variabilidad de

* Conferencia del 1er. Foro de Estadística Aplicada. UNAM. 26 Septiembre, 1986.

las plantas cultivadas, de los animales domésticos y su medio ambiente.

Durante muchos siglos de actividades agrícolas, se obtuvo un progreso importante tanto en las técnicas de producción como en el mejoramiento genético de las especies animales y vegetales útiles al hombre. Pero este progreso, importante como fué, no se obtuvo de manera sistemática, sino a lo largo de una gran serie de pequeños avances por "ensayo y error". Por ésto, fue muy importante para la historia de la agronomía y de la estadística, el hecho de que los directivos de la Estación Agrícola Experimental de Rothamsted invitaran en 1919, a R.A. Fisher (de formación matemática) a colaborar con ellos. Rothamsted tenía alrrededor de 80 años de realizar investigación agrícola, pero sin reconocer la necesidad del control de la variabilidad genética y ambiental de los cultivos. Fisher produjo toda una teoría para la experimentación agrícola, en la que podemos señalar como principales contribuciones:

1.- El uso de los modelos estadísticos lineales, con su teoría de estimación y pruebas de hipótesis; tales como la distribución F, la prueba no paramétrica basada en la aleatorización.

2.- El uso de bloques como medio de controlar factores de confusión, es decir factores de variación ajenos al estudio, pero presentes de modo diferencial en las unidades experimentales.

3.- El uso de la aleatorización con un papel dual, por un lado un control adicional de factores de confusión y por el otro el imprimir le validez a las pruebas de significación estadística.

4.- El concepto de confusión entre factores y los experimentos factoriales con confusión.

5.- La aplicación de la teoría de grupos para obtener esquemas convenientes de confusión parcial y total.

6.- El uso de la covarianza, como medio para lograr durante el análisis, a posteriori del experimento, un control de otros posibles factores de confusión.

Todos estos desarrollos quedaron plasmados en los muchos artículos que Fisher escribiera y en sus conocidos libros "Statistical Methods for Research Workers, (1925)", "The design of Experiments, 1935" y con F. Yates como coautor "Statistical Tables for Biological, Agricultural and Medical Research Workers". Ya en este último se observa una extensión que aun continúa hasta nuestros días, de los conceptos desarrollados en agronomía a otras áreas como la biología y medicina. Esta teoría con sus conceptos, leyes y metodología fue refinada y aplicada a campos como la Ingeniería, Geografía, Psicología, Química, etc.; por muchos otros autores alumnos o continuadores de Fisher, tales como Snedecor, Cochram, G. Cox, D. Cox, Box, Hunter, etc.

Además en la biología y la genética, Fisher desarrolló técnicas de análisis multivariado, así como la base matemática para la genética de poblaciones o genética cuantitativa, la que ha tenido una enorme importancia en el desarrollo de variedades de plantas y animales más productivos y mejor adaptados a sus condiciones de producción.

Como acertadamente señala Aranda (1982), los trabajos de Fisher, por supuesto, no arrancaron de la nada, había ya antecedentes importantes como los recientes desarrollos de Gosset o "Student" sobre el arreglo sistemático de experimentos con cereales, en 1908. Desde luego a su vez "Student" tenía otros antecesores en las ideas de experimentación agrícola, tales como las de A. Young que publicó en 1771 un tratado llarado "A course of Experimental Agriculture", en el que ya planteaba la necesidad de experimentos comparativos y la consideración de la variación climática y del suelo. Otro antecedente posterior fue J. Johnston con su libro "Experimental Agriculture" (1849), que también señalaba el control de la variación del suelo y la necesidad de evaluar la variación en los resultados (lo que "Student" llamó después el error probable, luego modificado a ErrorEstandar).

Desde principios del siglo se han realizado los llamados "Experimentos en blanco" o "ensayos de uniformidad", para estudiar la variabilidad geográfica en los campos agrícolas. Estos estudios se con

ducen en un terreno tratado tan uniformemente como sea posible, con la misma variedad y prácticas de cultivo. Se miden las respuestas (observaciones del rendimiento) de pequeñas parcelas, por separado para cada unidad junto con su localización geográfica dada por las coordenadas de la parcela. Se pensó que la variación presente obedecía a los cambios de fertilidad del suelo, sin embargo, también representa cambios en incidencia de plagas y enfermedades, vientos dominantes, y otros factores ambientales que producen efectos sitemáticos sobre el terreno cultivado. Con esos datos se contruyen mapas de fertilidad o de productividad geográfica, al suponer gradientes lineales de productividad entre los centros de parcelas contiguas. En los trabajos de Cochran (1937), Pearce (1953) y Méndez (1970) se presentan catálogos de ensayos de uniformidad.

En 1915, Harris utilizó coeficientes de correlación para medir la heterogeneidad de las producciones de parcelas vecinas, sin embargo, Smith en 1938 presentó una ley empírica que relaciona el tamaño de las parcelas y su variabilidad, al mismo tiempo que rechaza la utilidad de los coeficientes de correlación. Este autor utiliza el coeficiente de heterogeneidad b, para medir la variación geográfica. Su ley es $V(x) = V(1)/X^b$ donde $V(x)$ es la varianza de parcelas de tamaño X. En Méndez (1970) se presentan los métodos de Richey, Papadakis y Hoyle-Baker (1961) que proponen métodos alternativos a los de Fisher, para el diseño y análisis de experimentos agrícolas. Estos se basan en promedios móviles, índices de fertilidad como covariables y un método no paramétrico sobre "islas de variación", respectivamente.

Como vemos los métodos de Fisher, fueron muy poderosos e influenciaron muchísimos desarrollos posteriores, pero como todo modelo hacen supuestos sobre la realidad, que nunca se cumplen cabalmente, por lo que sacrifican precisión en aras de la sencillez conceptual y de análisis. Pero desde su origen prácticamente, ya había propuestas alternativas.

III Crítica al Concepto de Bloque en Agronomía y sus Alternativas

Las bases que Fisher estableció para los experimentos Agrícolas se extendieron prácticamente a cualquier actividad ajena a la agronomía, sin embargo, comentaré aquí dos críticas que sus modelos de bloques han recibido. En este apartado discuto lo relativo a la falla que surge al suponer que las parcelas dentro de un bloque son homogéneas. En el apartado siguiente discuto la falla al no considerar que los propios bloques no tienen repeticiones independientes.

Como ha quedado claramente demostrado en todos los ensayos de uniformidad, los patrones o tendencia de variación geográfica no ocurren en conjuntos de parcelas con igual productividad, en arreglos geométricos, que puedan coincidir con bloques. Además de que al iniciar un experimento los patrones de variación son desconocidos, lo que dificulta la ubicación adecuada de los bloques. En Méndez (1970) primero se encuentran modelos de funciones de tendencia en los ensayos de uniformidad. Se postuló el modelo $Y_{\alpha\lambda}=T_{\alpha\lambda}+E_{\alpha\lambda}$ donde $Y_{\alpha\lambda}$ es el rendimiento en la parcela con coordenadas α y λ, $T_{\alpha\lambda}$ es la función de tendencia geográfica de las producciones en la localidad α,λ y $E_{\alpha\lambda}$ un error aleatorio no asociado a su localización. Se ajustaron para la función de tendencia modelos polinomiales, de series de Fourier y polinomiales inversos. Con esta información, para un experimento agrícola se postula que además de los términos anteriores se tendrá un efecto del tratamiento j dado por M_j así el modelo resulta $Y_{ij}(\alpha\lambda) = M_j+T_{\alpha\lambda}+E_{ij}(\alpha\lambda)$. Se usan ahora doble subíndices para señalar repetición, i, tratamiento, j y posición geográfica α,λ. Se comparó mediante simulación este modelo con el usual de Fisher, resultando que el modelo de bloques de Fisher sobreestima la varianza de los errores y produce pruebas con menor potencia que el modelo propuesto, llamado "Tendencia sobre Residuos". También se encontró que el método de Papadakis resultó mejor que el diseño en bloques completos al azar. Estos resultados han sido usados para análisis de experimentos agrícolas en Estados Unidos y en México. En el paquete SAS (Statistical Analysis System) se incorporó una rutina para análisis de tendencia geográfica.

Otro uso de estas ideas fue el determinar potencialidad productiva de variedades de cacao en Tabasco, al eliminar la tendencia geográfica (a nivel de todo el estado) para cada predio y estimar así el efecto de variedad eliminando efectos ambientales; esto fue la tesis de Maestría de J. Somellera en el extinto Colegio Superior de Agricultura Tropical, 1984. Se han iniciado ya trabajos que incorporan conceptos de procesos estocásticos en un plano para el ajuste de la función de tendencia.

IV Error de Restricción

Durante 50 años se utilizaron las ideas de bloques en todos los campos de investigación sin considerar un aspecto importante, se asigna en los modelos un efecto de bloque igual conceptualmente al de los tratamientos y sin embargo en la realidad el bloque no se aleatoriza (salvo la etiqueta que se le asigna) y los tratamientos si se asignan al azar a las unidades. Esto establece una asimetría que no fue reconocida por Fisher, ni sus inumerables seguidores, durante esos años y aun hoy es poco conocida en muchas instituciones. No fue sino hasta 1970 cuando V.L. Anderson publicó un artículo sobre este aspecto que llamó el "error de restricción", posteriormente en 1974, junto con McLean publicó un libro de diseño de experimentos, en el que aplica el concepto a muchos diseños experimentales. El no considerar este error de restricción puede llevar a cometer errores en las inferencias, como demostró el mismo Anderson.

Si se tiene un grupo de unidades experimentales donde se supone que ocurre un efecto β en <u>todas</u> las unidades <u>simultáneamente</u>, esto no equivale a la ocurrencia independiente de β, en cada unidad experimental. Los aspectos aleatorios que caracterizan ese grupo de unidades quedarán confundidos con β_j. Así este concepto distingue entre dos situaciones experimentales que frecuentemente se confunden en la práctica:

Fig. 1

Asignación aleatoria de combinaciones de niveles de factores a las unidades

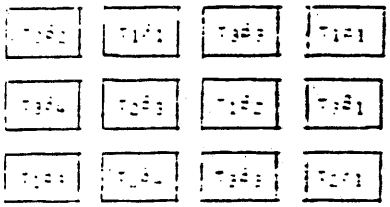
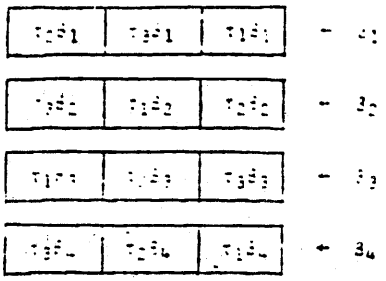


Fig. 2

Asignación aleatoria de los tratamientos dentro de cada bloque



t = 3 niveles de tratamientos.

b = 4 niveles de bloques.

Tradicionalmente ambas situaciones se analizan con el mismo modelo, que supone ausencia de interacción entre bloques y tratamientos:

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{(ij)} \qquad (1)$$

donde:

μ efecto de media general

τ_i efecto de tratamiento i

β_j efecto de bloque j. En general i=1,...,t, j=1,...,b

$\varepsilon_{(ij)}$ error aleatorio

Claramente las situaciones de Fig. 1 y de Fig. 2 no son iguales, por lo tanto el modelo (1) no puede representar correctamente ambas situaciones. El error de restricción permite adecuadamente dis-

tinguir entre esas dos situaciones. Así el modelo (1) es adecuado para la situación de Fig. 1, pero no lo es para la de Fig. 2. Para este último caso se usa el modelo (Anderson).

$$Y_{ijlk} = \mu - \tau_i + \beta_j + \delta_{l(j)} + \varepsilon_{k(ijl)} \qquad 2)$$

$$\begin{matrix} i=1,\ldots,t & l=1 \\ j=1,\ldots,b & k=1 \end{matrix}$$

Donde $\delta_{l(j)}$ es el error de restricción de bloques, y se introduce en el modelo para representar las características particulares y aleatorias de ese conjunto de unidades que forma el bloque j; tales como errores de medición, condiciones ambientales comunes, etc. Se considera a $\delta_{l(j)}$ como aleatorio ya que las características peculiares de ese bloque no son reproducibles a voluntad por el investigador.

En otros diseños experimentales cuando hay grupos de unidades con características comunes y dentro del grupo se asignan algunos factores, se usará un error de restricción para los aspectos comunes de ese grupo de unidades.

El error de restricción es un efecto confundido (no se puede separar del) con el efecto de bloque o grupo de unidades.

El error de restricción no es estimable, ni las combinaciones de esos errores, porque no hay disponibles grados de libertad (g. l.) para la estimación. Sin embargo, al incluirlo en el modelo y obtener esperanzas de cuadrados medios E(CM), se puede notar qué efectos se pueden probar mediante razones F.

La Ho: $\tau_1 = \tau_2 = \ldots = \tau_t$ se puede probar mediante F = CMtrat/CM error aleatorio. Sin embargo, la Ho: $\beta_1 = \beta_2 = \ldots = \beta_b$ no se puede probar. Esto concuerda con la intuición ya que no hay observaciones repetidas independientes de cada β_j.

En los experimentos con animales mayores, al igual que con las plantas perenes se pueden tener unidades experimentales grandes y costosas con mucha variabilidad. No debe caerse en el error de manejar grupos de unidades en forma conjunta sin asignar un error de restricción.

Este concepto de error de restricción permite darle mucha claridad y lógica a los diseños de parcelas divididas, también propuestos por Fisher y Yates, como "experimentos complejos". Así, surge de manera natural, basada en las esperanzas de cuadrados medios, la recomendación de que cada tamaño de parcela tiene un error experimental diferente del resto.

Otro uso sumamente importante del error de restricción, es que permite introducir en el análisis una diferencia clara entre un experimento y un pseudoexperimento donde no se aleatorizaron los factores en estudio. Esto último cobra importancia capital en la investigación social.

V Otros Desarrollos

Podemos afirmar que en la agronomía como en otras ciencias, se dió el mismo tipo de estimulación recíproca con la estadística; en el sentido de que su crecimiento fue simultáneo y estimulado mutuamente. Así el desarrollo de la teoría de la experimentación estimuló la agronomía y ésta a su vez a la estadística. Pero este proceso no se limitó al tópico del diseño de experimentos, también se dió en temas como análisis multivariado y muestreo. Por ejemplo el análisis discriminante, ¡También iniciado por Fisher¡, se planteó inicialmente con fines de taxonomía de plantas. El muestreo se aplicó y se desarrolló en parte, por la necesidad de muestrear plantas, suelos e insectos, para llegar a decisiones sobre la necesidad de fertilización o del control de hierbas, plagas o enfermedades.

La necesidad de predecir las cosechas y las condiciones de

lluvia, temperatura, granizo, heladas, etc.; se apoyó y estimuló el desarrollo de las series de tiempo y procesos estocásticos.

El problema de estudiar la producción de plantas que fructifican en varias épocas como el tomate o los frutales, generó los modelos de "observaciones repetidas", ahora muy usados en las ciencias de la conducta humana. Desde luego, como paso intermedio se aplicaron y mejoraron en los experimentos con animales. Toda esta área de la investigación fue mejorada mucho al desarrollarse concomitantemente las técnicas de análisis de varianza multivariado y las de computación.

En el aspecto de la búsqueda de las combinaciones óptimas de factores de la producción, tales como cantidad de nitrógeno, fósforo y potasio agregado al suelo, densidad de siembra, dósis de plaguicidas, etc; se desarrolló la "Metodología de Superficie de Respuesta". De estas aplicaciones, pronto se pasó a la ingeniería industrial, donde fue mejorada mucho por Box, Wilson y Hunter entre otros. Con estos desarrollos de nuevo aplicados a la agronomía, al agregar conceptos económicos, se generó la técnica de funciones de producción agrícola, tan brillantemente expuesta por Heady y Dillon en su libro de 1961, con ese mismo nombre. Estas técnicas consisten en el uso de modelos de regresión múltiple con funciones de las variables independientes como potencias, logaritmos y funciones trigonométricas, para lograr representaciones adecuadas de la influencia de los insumos en los procesos productivos. Después, con ayuda de la economía, encontrar los puntos que optimizan económicamente esos procesos.

Es constante la interrelación mutua entre la agronomía como ciencia que estudia los procesos de producción agropecuaria y forestal, por un lado; y la estadística como herramienta para la obtención y análisis de información en presencia de variabilidad, por el otro. Así, se realiza investigación encaminada a buscar nuevas aplicaciones agronómicas de los métodos estadísticos, la modificación de los métodos actuales para que los supuestos que hacen los modelos estén más cercanos a la realidad y también la generación de métodos nuevos para

problemas nuevos o viejos que surgen al avanzar la agronomía.

La agricultura se ha definido como un proceso de producción determinado histórica y socialmente, que se caracteriza por que el hombre aplica su fuerza de trabajo, sus conocimientos y habilidades a través de medios e instrumentos, para el aprovechamiento y transformación del medio físico y biológico, con el fin de obtener bienes vegetales y animales. La investicación agrícola se ha enfocado principalmente a los objetos de trabajo: plantas, animales, fertilizantes, pesticidas, etc. y a las condiciones naturales y artificiales como suelo, clima, predatores. Se han dejado de lado o se han abordado sólo marginalmente los aspectos de relaciones sociales y fuerza de trabajo. Para poder estudiar estos últimos aspectos, es necesario tomar la metodología de investigación desarrollada en ciencias sociales, para que se aplique al estudio de la agricultura. Esto implica que los métodos estadísticos para los estudios observacionales deban de usarse con mucho más frecuencia en los estudios sobre agricultura.

VI Bibliografía

- Aranda, O.F. "Diseños Experimentales: algunos comentarios sobre su desarrollo y fundamentos". Comunicaciones Técnicas. Serie Azul, No. 60, IIMAS, UNAM, 1982, p.13.
- Neyman, J. "R. A. Fisher (1890-1962): An appreciation". Science, 156, 1967, 1456-1460.
- Heady, E.O. y Dillon, J. L. "Agricultural Production Functions". Ames, Iowa State University Press, 1961.
- Fisher, R.A. "Statistical Methods for Research Workers", Edinburgo, Oliver and Boyd. 1925.
- Fisher, R. A. "The Design of Experiments", Edinburgo, Oliver and Boyd, 1935.
- Méndez, R. I. "Study of Uniformity trials and sixs proposals as alternatives to blocking for the design and analysis of field experiments", Ph.D. Thesis.
  Institute of Statistics Mimeo Series No. 696, North Carolina State University. Raleigh, N.C., 1970, p. 241.
- Cochran, W.G. "A catalogue of Uniformity Trials". J. Roy. Stat. Soc. 4:233-253.1937.
- Pearce, S.C. "Field Experimentation with fruit trees and other perennial plants". Technical communication No. 23 of the Commonwealth - Bureau of Horticulture and Plantation Crops. East Malling. England. 1953.
- Anderson V. L. "Restriction errors for linear models" (An aid to - develop models for designed experiments). Biometrics 25:255-268, 1970.
- Anderson V. L. y R. A. Mc. Lean "Design of Experiments: A realistic approach", New York, Marcel Dekker, 1974.

UNA APLICACION DE MODELOS LOGLINEALES A LA BACTERIOLOGIA MEDICA

Raúl Rueda

Universidad Autónoma Metropolitana-Iztapalapa

INTRODUCCION

La posibilidad de aislar bacterias de diferentes muestras, es de gran ayuda al médico en el diagnóstico de enfermedades causadas por bacterias anaerobias, ya sea que estas bacterias actúen en forma aislada o bien asociadas a otros microorganismos. Al aislar a estas bacterias, puede estudiarse que tan susceptibles resultan ser a ciertos antimicrobianos anaerobios de interés clínico, de manera que pueda darse un tratamiento adecuado a los pacientes infectados con ellas.

Sin embargo, existen diferentes métodos para realizar estos estudios: algunos son sencillos y baratos, pero poco confiables; mientras que otros son mas exactos, sólo que son costosos y complicados. Por esta razón, personal del laboratorio de Bacteriología Médica del Centro Hospitalario "20 de Noviembre" del ISSSTE, decidió efectuar un experimento donde se probarían dos métodos que eran fáciles de usar y otros cuatro que, se había demostrado ya, eran confiables. El objetivo principal era determinar si los métodos influían en la detección de la susceptibilidad de los microorganismos a los antimicrobianos. La idea era probar que los dos métodos nuevos eran igualmente confiables que los otros cuatro.

El experimento consistió en lo siguiente: se aislaron 50 cepas de microorganismos anaeróbicos de enfermos infectados. Estas cepas fueron agrupadas en nueve grupos, usando como criterio el de familia. Cada grupo fué combinado, mediante inoculación, con siete tipos de antimicrobianos, usando los seis métodos. Las diferencias básicas entre estos métodos fueron: el caldo en donde el microorganismo fué cultivado, el orden de inoculación y el preparado en donde se tenía al antimicrobiano. La susceptibilidad del microorganismo al antimicrobiano fué medida en términos de la turbidez de la solución.

Desde el punto de vista estadístico, se tienen cuatro variables: tres

controladas en el experimento (microorganismo, antimicrobiano y método) y una cuarta que es el resultado del mismo y por tanto es una variable de respuesta (susceptibilidad).

Esta información fué arreglada en una tabla de contingencia con cuatro criterios de clasificación:

Microorganismo	Antimicrobiano	Método	Respuesta	
		1		
		2	n_{i111}	n_{i112}
		.	n_{i121}	n_{i122}
	1	.		
		6	.	
			.	
			.	
		1		
		2		
		.		
	2	.		
		6		
	.			
	.			
	.			
		1	n_{i711}	n_{i712}
		2	n_{i721}	$n_{i72.}$
		.	.	
	7	.	.	
		6		
			n_{i761}	n_{i762}

en donde n_{ijk_1} es el total de microorganismos del tipo i que resultaron susceptibles al antimicrobiano j inoculado con el método k; mientras

que n_{ijk_2} es el total de microorganismos resistentes, con las mismas características.

En vista que la variable respuesta está medida en una escala nominal, se decidió ajustar un método logístico, que permitiría estudiar la influencia de cada método en la susceptibilidad de los microorganismos a los antimicrobianos utilizados.

A "grosso modo", un modelo logístico es el equivalente a un modelo lineal de análisis de varianza, sólo que para los datos categóricos.

El modelo saturado -el que contiene todos los términos- en este caso, está dado por

$$(1) \quad \lambda_{ijk} = \omega + \omega_1(i) + \omega_2(j) + \omega_3(k) + \omega_{12}(ij) + \omega_{13}(ik) + \omega_{23}(jk) + \omega_{123}(ijk)$$

en donde $\lambda_{ijk} = \log_e \frac{\theta_{ijk}}{1 - \theta_{ijk}}$: probabilidad de que un microorganismo i sea susceptible al antimicrobiano j al combinarse por el método k.

es decir, $_{ijk}$ representan los log-momios a favor de la susceptibilidad; ω_I efecto de las variables I sobre la respuesta.

Por ejemplo, $\omega_{12(ij)}$ representa el efecto de los niveles i,j de las variables 1,2 en la variable respuesta.

Existen diferentes caminos para analizar estos modelos

i.- Métodos lineales generalizados (e.g. Nelder & Wedderburn, 1972)
ii.- Mínimos cuadrados ponderados (e.g. Cox, 1970).
iii.- Modelos loglineales (Bishop, 1969; Fienberg, 1977; Goodman, 1971).

En este caso se usaron modelos loglineales. Esta decisión fué debida a la capacidad de cómputo instalada que pudo usarse.

Un modelo loglineal, a grandes rasgos, es útil para describir la relación estructural entre todas las variables que componen una tabla de contingencia. En el caso de que exista una variable respuesta, el modelo logístico puede ser obtenido fácilmente a partir del modelo loglineal.

Para este problema, el modelo loglineal saturado es:

$$\log_e F_{ijkl} = u + u_1(i) + u_2(j) + u_3(k) + u_4(l) + u_{12}(ij) + u_{13}(ik) + u_{14}(il) + u_{23}(jk) + u_{24}(jl) + u_{34}(kl) + u_{123}(ijk) + u_{124}(ijl) + u_{134}(ikl) + u_{234}(jkl) + u_{1234}(ijkl) \qquad (2)$$

en donde F_{ijkl} es la frecuencia esperada en la celda ijkl de la tabla y los términos u_l tienen una interpretación similar al caso anterior. A partir del modelo especificado con (2) puede obtenerse el modelo loglineal (1) haciendo

$$\lambda_{ijk} = \log_e F_{ijk_1} - \log_e F_{ijk_2}$$

En el caso en que un modelo logístico sea ajustado a través de un modelo loglineal, Bishop (1969) menciona que el término que describe la interacción mas grande entre las variables controladas debe incluirse en el modelo para producir estimaciones estables de los términos ω_l.

El tipo de hipótesis que se prueban en estos modelos son de la forma

$H_l: u_l = 0$ en el caso del modelo loglineal, o

$H_l: \omega_l = 0$ en el caso del modelo logístico.

En donde I denota a algún conjunto de índices.

El criterio usado está basado en la divergencia logarítmica de Kullback & Leibler (1951)

$$G = 2 \sum OBS \log_e \frac{OBS}{ESP}$$

en donde la suma es sobre todas las celdas de la tabla y ESP es el valor esperado en cada celda bajo el modelo especificado en la hipótesis. Si H es cierta, G se distribuye asintóticamente como una χ^2 con grados de libertad igual a la diferencia entre el número de celdas y el número de parámetros estimados en el modelo correspondiente*.

En general se tiene un problema de selección de modelos: encontrar un modelo que contenga el menor número de parámetros posible, pero al mismo tiempo que explique en forma razonable el conjunto de datos que se tiene. Para resolverlo se supone que los modelos que se ajustan son jerárquicos, lo cual significa que si un término se incluye en el modelo, todos los términos de orden menor que contengan a los índices del primero, deben ser incluidos; además, sólo se consideran modelos anidados.

La estadística de prueba que se usa, suponiendo que el modelo dos está anidado en el uno, es

$$G(2|1) = 2 \sum OBS \log_e \frac{ESP_2}{ESP_1}$$

que, si H es verdadera, se distribuye como χ^2 con grados de libertad igual a la diferencia de los grados de libertad asociados a cada modelo.

Bajo estos criterios, usando el modelo loglineal saturado (2) se encontró que el modelo

---

* Suponiendo que ninguna celda tiene asociado un valor de cero. Si existen celdas con cero, deben ajustarse los grados de libertad (ver e.g. Bishop, Fienberg & Holland, 1975).

$$\log_e F_{ijkl} = u + u_1 + u_2 + u_3 + u_4 + u_{12} + u_{13} + u_{14} + u_{23} + u_{24} + u_{123} + u_{124}$$

explicaba razonablemente la estructura de la tabla y contenía el menor número de parámetros.

El modelo establece que las variables método y respuesta (3 y 4) son independientes dadas las variables microorganismo y antimicrobiano (1 y 2).

El modelo logístico asociado resulta ser

$$\lambda_{ijk} = \omega + \omega_1 + \omega_2 + \omega_{12} \qquad (3)$$

que especifica que la respuesta depende del efecto combinado del microorganismo y antimicrobiano, pero no del método.

En conclusión, el modelo logístico ajustado permite decir, con los datos disponibles, que todos los métodos tienen el mismo efecto en la detección de la susceptibilidad; de hecho, posteriormente se efectuó otro experimento considerando más variables y otras combinaciones microorganismo-antimicrobiano, obteniéndose resultados similares. Por último, en el modelo (3), los parámetros pueden estimarse para tener una idea de que combinaciones microorganismo-antimicrobiano producen con mayor frecuencia una respuesta "susceptible" y cuales "resistente", aún mas, cada variable puede ser estudiada en cada uno de los niveles de la otra y determinar sus efectos en la respuesta. (Véase Rueda, 1984).

R E F E R E N C I A S

BISHOP, Y.M.M. (1969). Full contingency tables, logits and split contingency. Biometrics 25, 383-399.

BISHOP, Y.M.M; FIENBERG, S.E. & HOLLAND, P.W. (1975). Discrete Multivariate Analysis. MIT Press.

COX, D.R. (1970). Analysis of Binary Data. Methuen, Londres.

FIENBERG, S.E. (1977). The analysis of Cross Classified Categorical Data. MIT Press.

GOODMAN, L.A. (1971). The analysis of multidimensional contingency tables: stepwise procedure and direct estimation methods for building models for multiple classification. Technometrics 13, 33-61.

KULLBACK, S. & LEIBLER, R.A. (1951). On information and sufficiency. Amer. Math. Stat. 22, 79-86.

RUEDA, R. (1984). Una aplicación de modelos loglineales a un estudio comparativo de seis métodos para la determinación de la susceptibilidad a los antimicrobianos y gérmenes anaerobios. Publicaciones del Departamento de Matemáticas, U.A.M.I.

UNA INTRODUCCION AL ANALISIS DE SUPERVIVENCIA

M. en C. Belem Trejo Valdivia, IIMAS-UNAM

RESUMEN

Se presenta un panorama del análisis más general de datos de supervivencia, haciendo énfasis en la importancia de la censura en dichos datos. Este panorama abarca tanto la parte paramétrica como la no-paramétrica de datos provenientes de una población homogénea o heterogénea. En la parte no-paramétrica se presenta con mayor detalle el estimador Kaplan-Meier de la función de supervivencia.

I.- INTRODUCCION.

Los llamados datos de supervivencia [1] , son aquellos que se generan al estudiar, en una población de interés, el tiempo que transcurre entre la ocurrencia de dos eventos especiales. El primer evento determina la entrada de un individuo al estudio, este primer evento se conoce como "entrada al estudio" o "nacimiento de dicho individuo". Se supone que se cuenta con n individuos que inician el estudio (no necesariamente en el mismo momento) y se les sigue en el tiempo hasta la ocurrencia del segundo evento en cada uno de ellos, dicho evento recibe el nombre de "falla" o "muerte".

Como puede verse, la definición es bastante amplia con lo cual este tipo de datos pueden situarse en diferentes áreas.

En el caso de poder determinar el tiempo de falla en cada individuo de la muestra, es posible utilizar los métodos tradicionales estadísticos para estudiar el comportamiento del tiempo de supervivencia, sin embargo, al tener un estudio longitudinal (pues existe un seguimiento en el tiempo) surge la posibilidad de CENSURA, es decir, el no poder observar a los individuos hasta la ocurrencia del segundo evento de interés.

---

1. Los nombres utilizados en este contexto se derivan del hecho que los primeros estudios de este tipo fueron los de nacimiento-muerte en poblaciones humanas.

La censura puede deberse a varias razones lo que produce que se distingan entre varios tipos:

1.- Censura tipo I.
Este tipo de censura se puede presentar cuando se fija el tiempo de terminación del estudio, llamado tiempo fijo de censura, independientemente de la muestra, esto es, el estudio puede haber terminado mientras que algunos individuos aún están vivos.

2.- Censura tipo II.
La situación clásica en donde se tiene este tipo de censura es cuando se decida concluir el estudio después de observar la r-ésima falla ($r < n$), de donde se tendrá que algunos individuos aún están vivos.

3.- Censura aleatoria.
Hay dos situaciones que se contemplan en este rubro:
- La falla puede presentarse en algunos individuos por causas que no son de interés para el estudio.
- Algunos individuos en los que no se ha presentado la falla pueden retirarse del estudio sin que este haya concluído aún.

Sin embargo, esta distinción no es considerada dentro del análisis general de los datos, ya que las observaciones de una muestra son tratadas simplemente como no censuradas y censuradas.

Por lo anterior, es necesario desarrollar procedimientos especiales que consideren la presencia de censura en la muestra, para poder analizar datos de supervivencia.

Los siguientes ejemplos demuestran que este tipo de datos pueden presentarse en muy diversas áreas:

1.- Control de Calidad.
En una inspección sobre la calidad de cierto tipo de aparatos eléctricos, se decide poner a trabajar en condiciones normales un número fijo de éstos y estudiar el tiempo que tardan en presentar la primera falla. Por lo tanto, el evento de interés es la presencia de cualquier falla en estos aparatos y la variable respuesta es el tiempo que transcurre desde el momento en que es puesto a funcionar un aparato hasta que presenta la primera falla.

2.- Estudios de fecundidad.
Dentro de esa área, una variable de interés es la edad de la mujer al momento de tener su primer hijo. Entonces, la variable a estudiar es el tiempo que transcurre desde el momento que una mujer es fértil (entrada al estudio) hasta que nace su primer hijo vivo (falla). En este caso, los valores censurados pueden deberse, entre otras cosas, a mujeres que:
-durante el estudio mueren por causas diferentes al alumbramiento del primer hijo vivo.
-abandonan el estudio por cambiar de residencia y no es posible localizarlas posteriormente.

3.- Educación.
En un análisis sobre el desarrollo de algún proyecto docente (licenciatura, especialización, maestría o doctorado) se decide estudiar el tiempo que tarda un alumno de este proyecto en obtener el título correspondiente. El segundo evento de interés en este caso es la obtención del título y la variable respuesta es el tiempo que transcurre desde que el alumno termina el 100% de los créditos hasta el momento en que obtiene el título. Entre las posibles causas de observaciones censuradas están:
- El estudio termine y haya alumnos que habiendo terminado los créditos correspondientes no hayan obtenido el título aún.
- Alumnos que abandonan el estudio ya sea por muerte, cambio de residencia, renuncia a obtener el título, etc.

El tipo de análisis que se lleva a cabo depende fuertemente de las condiciones que se tengan al inicio del estudio. Si todos los individuos que inician el estudio tienen características similares y no existen factores propios de ellos, que alteren de alguna forma la respuesta, se considera que se tiene una población homogénea, en cualquier otro caso, se considera que la población es heterogénea. En cualquier caso, T será una variable aleatoria no negativa que representa el tiempo de falla o tiempo de supervivencia.

II.- ANALISIS EN POBLACIONES HOMOGENEAS

En el caso de una población homogénea, el comportamiento de T, se lleva a cabo en términos de tres diferentes funciones relacionadas entre sí, cada una de ellas con cierta utilidad y aplicación dentro del análisis de supervivencia. Dichas funciones son: La función de supervivencia, la función de densidad de probabilidad y la función de riesgo.

Función de supervivencia.-

Esta función se define como la probabilidad de que un individuo sobreviva hasta el tiempo t, esto es, la probabilidad de que T sea mayor o igual que t, en donde t es una valor positivo dado. Es decir [2],

$$S(t) = P[T \geq t] \qquad \text{con} \qquad 0 < t < \infty$$

la cual es una función monótona no-creciente y contínua por la izquierda, con $S(0) = 1$ y $\lim_{t \to \infty} S(t) = 0$.

Aunque esta función es esencialmente el complemento a 1 de la función de distribución de T, en este contexto es mejor trabajar con S(t) por cuestiones de interpretación.

Función de densidad de probabilidad.-

Esta función se define de la manera tradicional, es decir,

$$f(t) = \lim_{h \to 0^+} \left\{ \frac{P[t \leq T < t + h]}{h} \right\}, \quad 0 < t < \infty$$

Es fácil demostar a partir de la definición que esta función y la de supervivencia están relacionadas a través de las expresiones:

$$f(t) = -\frac{d}{dt} S(t)$$

$$S(t) = \int_t^{\infty} f(u)du$$

---

2. Por simplicidad sólo se presentan las expresiones para el caso en que T es contínua, las correspondientes al caso discreto y/o mixto son directas pues las definiciones se dan de manera general.

Función de riesgo.-

Esta función mide la tasa instantánea de falla en T=t condicionada a la supervivencia al tiempo t (en demografía es llamada la fuerza de mortalidad), es decir,

$$\lambda(t) = \lim_{h \to 0^+} \left[\frac{P[t \leq T < t + h | T \geq t]}{h} \right]$$

de donde la relación con las dos funciones anteriores se expresa por:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \operatorname{Ln} S(t)$$

$$S(t) = e^{-\int_0^t \lambda(u)du}$$

$$f(t) = \lambda(t)\, e^{-\int_0^t \lambda(u)du}$$

Dadas las relaciones entre estas tres funciones, bastará estimar, ajustar o modelar una de ellas para tener idea del comportamiento de las otras, con lo que globalmente se produce un amplio conocimiento acerca de T.

La función que se modela depende del enfoque con que se lleve a cabo el ajuste, dicho enfoque, idealmente, está determinado por el conocimiento (o experiencia) que se tenga del problema en estudio. El enfoque puede ser paramétrico o no-paramétrico.

MODELOS PARAMETRICOS PARA TIEMPO DE FALLA.

Se han propuesto algunos modelos paramétricos sobre la distribución de T, entre los cuales se encuentran los siguientes:

1.- Modelo Exponencial.
2.- Modelo Weibull.
3.- Modelo Log-normal.
4.- Modelo Gamma.

Los más comúnmente utilizados son los dos primeros, los cuales se enfocan a modelar la función de riesgo. Los otros dos (y algunos mas que no se han mencionado) especifican el comportamiento de la función de densidad de probabilidad. En este trabajo se presentarán brevemente los dos primeros modelos.

Modelo Exponencial.-

Este modelo es el mas simple e importante en estudios de supervivencia ya que juega un papel análogo al de la distribución normal en otras areas de estadística. Este modelo postula que la tasa instantánea de falla es independiente de t, es decir, la función de riesgo es constante sobre el rango de T

$$\lambda(t) = \lambda > 0 , \qquad 0 < t < \infty$$

Las funciones de supervivencia y de densidad de probabilidad (de donde se deriva el nombre del modelo) están dadas, respectivamente por

$$S(t) = \begin{cases} e^{-\lambda t} & t \geq 0 \\ 1 & t < 0 \end{cases}$$

y

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

Modelo Weibull.-

Este modelo es una generalización del modelo exponencial, puede utilizarse para modelar el comportamiento de la supervivencia en una población con función de riesgo creciente, decreciente o constante. Por lo tanto se supone que el riesgo depende fuertemente del tiempo, lo cual se refleja a través de dos paramétros positivos, λ un parámetro de escala y p un parámetro de forma, expresándose

$$\lambda(t) = \lambda p(\lambda t)^{p-1} \qquad 0 < t < \infty.$$

Esta función de riesgo es monótona decreciente para $p<1$ y monótona creciente para $p>1$. Se reduce al modelo exponencial cuando $p=1$.

La función de supervivencia es

$$S(t) = \begin{cases} e^{-(\lambda t)^p} & t \geq 0 \\ 1 & t < 0 \end{cases}$$

y la función de densidad de probabilidad es

$$f(t) = \begin{cases} \lambda p(\lambda t)^{p-1} e^{-(\lambda t)^p} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

ANALISIS NO-PARAMETRICO DEL TIEMPO DE FALLA.

En ese enfoque, el análisis se ha centrado principalmente en la estimación de la función de supervivencia, entre los diferentes estimadores que se han propuesto, los mas comunes son:

La funcion de supervivencia empírica.-

Esta función se define de la siguiente manera

$$S_n(t) = \frac{\text{Número de valores muestrales} \geq t}{n}$$

y sólo es útil en el caso de tener una muestra no censurada de tamaño n de tiempos de falla.

La tabla de vida.-

Es el más antiguo de los estimadores de la función de supervivencia. Esta tabla es un resumen de los datos de supervivencia agrupados en intervalos, es decir, presenta el número de fallas y censuras en cada intervalo. De manera general se pude decir que esta forma de estimación es muy pobre pues se pierde la información sobre los tiempos observados, además de depender fuertemente de la elección de los intervalos.

La tabla de vida modificada.-

Este estimador es una simple mejora al anterior, pues además de considerar los datos en una tabla de vida, toma en cuenta, los tiempos de falla y los tiempos de censura de todos los individuos observados. Aunque produce un estimador contínuo en t, depende también de la elección de los intervalos.

El estimador Kaplan-Meier.-

Es también llamado el estimador del producto-límite, fue propuesto en 1958 por Kaplan y Meier como un estimador simple de la función de supervivencia para el cual, no es necesario agrupar los tiempos de falla y de censura en intervalos arbitrarios. Este estimador es el que últimamente ha sido más estudiado, puesto que, se ha encontrado que posee algunas características deseables estadísticamente y lo único que supone es que se conocen los tiempos de falla. Por lo anterior, este estimador se presenta con más detalle; su construcción es como sigue:

Sean $t_1 < t_2 < \ldots < t_k$ los tiempos de falla observados en una muestra de tamaño n de una población homogénea con función de supervivencia S(t).

Supongamos que al tiempo t_j $(j=1,\ldots,k)$ se registran d_j fallas y en el intervalo $[t_j, t_{j+1})$ se presentan c_j censuras con $j = 0,\ldots,k$ en donde $t_0=0$ y $t_{k+1} = \infty$. Además, sea n_j el número de individuos expuestos al riesgo inmediatamente antes de t_j, es decir,

$$n_j = \sum_{x=j}^{k} (d_x + c_x)$$

Entonces el estimador Kaplan-Meier es

$$\hat{S}(t) = \prod_{j | t_j < t} \left(1 - \frac{d_j}{n_j}\right) \quad 0 \leq t < \infty$$

Este estimador es una función escalonada con saltos en las observaciones no censuradas. El tamaño del salto en cada observación no censurada es una función del tamaño de la muestra y el patrón de pérdida que se ha presentado antes de la falla. Un inconveniente de este estimador es que si el último valor observado corresponde a un tiempo de censura, esto es, $c_k > 0$, quedará indefinido el valor de $\hat{S}(t)$ para los t mayores que el valor censurado, algunos autores proponen definir $\hat{S}(t) = 0$ a partir del último valor observado independientemente de la naturaleaza de la última observación. Si en la muestra no hay observaciones censuradas (es decir, $k = n$), este estimador se reduce a la función de supervivencia empírica.

Como se dijo anteriormente, las propiedades de este estimador han sido estudiadas por varios autores, los cuales han mostrado entre otras cosas que:

- Es un estimador de tipo no-paramétrico.
- Es el estimador de máxima-verosimilitud generalizado.
- Es un estimador fuertemente consistente.
- Es asintóticamente normal.
- Visto como un proceso estocástico en t, converge débilmente a un proceso gaussiano.

III.-ANALISIS EN POBLACIONES HETEROGENEAS.

La no homogeneidad en la población, esto es, la discrepancia en el comportamiento del tiempo de falla entre los individuos de la población de estudio, se supone que se debe a la existencia de covariables $Z_1, Z_2, \ldots, Z_p$ que alteran la distribución de T. Estas covariables pueden ser propias de cada individuo y/o características de interés que son controladas por el investigador (ciertas condiciones experimentales, aplicación de algún tratamiento,etc.)

Los modelos paramétricos vistos anteriormente, pueden generalizarse de tal manera que se considere la información de los valores de las covariables de cada individuo de la muestra, para reflejar la dependencia de estas sobre el tiempo de falla.

Por lo tanto, supondremos que para cada individuo de la muestra se observarán el tiempo de falla o de censura y un vector $\underline{Z}' = (z_1, z_2, \ldots, z_p)$ de covariables. Lo mas común en este caso es modelar la función de riesgo vía un modelo aditivo o multiplicativo.

Si el modelo es aditivo, entonces se tendrá que

$$\lambda(t|\underline{Z}) = \lambda_1(t) + g(\underline{Z})$$

donde $\lambda_1(t)$ representa el riesgo común a todos los individuos. Este tipo de modelo es poco utilizado por cuestiones de interpretación.

Si el modelo es multiplicativo, entonces

$$\lambda(t|\underline{Z}) = \lambda_1(t)\, g(\underline{Z}) \ ,$$

de nuevo, $\lambda_1(t)$ representa el riesgo común a todos los individuos al tiempo t. La elección de g(.) puede depender del tipo de estudio que se está considerando, pero la elección mas comúnmente usada es $g(x)=\exp(x)$ suponiendo que la contribución

del vector de covariables es en forma lineal, de donde se obtiene el modelo de regresión

$$\lambda(t|\underline{Z}) = \lambda_1(t)\, e^{\underline{\beta}'\underline{Z}}$$

con $\underline{\beta}$ un vector de parámetros.

Este modelo es llamado el modelo de riesgos proporcionales y fue introducido por Cox (1972). En este caso, $\lambda_1(t)$ es una función arbitraria desconocida que da el riesgo al tiempo t para un individuo con condiciones estándar $\underline{z}=\underline{0}$.

Dentro de los modelos de riesgos proporcionales, el mas simple es aquel que supone $\lambda_1(t)$ constante a lo largo del tiempo; con esto se construye el llamado modelo exponencial de riesgos proporcionales

$$\lambda(t|\underline{Z}) = \lambda e^{\underline{\beta}'\underline{Z}}$$

De este modelo se obtiene que la función de densidad de probabilidad condicional de T dado $\underline{Z}$ es

$$f(t|\underline{Z}) = \lambda e^{-\underline{\beta}'\underline{Z}}\, e^{-\lambda t e^{\underline{\beta}'\underline{Z}}}$$

y la función de supervivencia condicional de T dado $\underline{Z}$ es

$$S(t|\underline{Z}) = e^{-\lambda t\, e^{\underline{\beta}'\underline{Z}}}$$

Una generalización del modelo anterior es el modelo Weibull de riesgos proporcionales en donde

$$\lambda(t|\underline{Z}) = p(\lambda t)^{p-1} e^{\underline{\beta}'\underline{Z}}$$

por lo que las expresiones para la función de densidad de probabilidad y la función de supervivencia condicionales de T dado $\underline{Z}$ son respectivamente

$$f(t|\underline{Z}) = p(\lambda t)^{p-1} e^{\underline{\beta}'\underline{Z}} - t(\lambda t)^{p-1} e^{\underline{\beta}'\underline{Z}}$$

y $$S(t|\underline{Z}) = e^{-t(\lambda t)^{p-1} e^{\underline{\beta}'\underline{Z}}}$$

De igual manera, se generalizan los modelos para poblaciones homogéneas, obteniéndose sus equivalentes para estas poblaciones heterogéneas. En esta parte se han propuesto técnicas de estimación para los parámentros de $\lambda(t|\underline{Z})$ ($\underline{\beta}$ y los asociados a $\lambda_1(t)$), entre las cuales se puede mencionar máxima-verosimilitud parcial.

BIBILIOGRAFIA.

1.- Breslow,N.E. (1975)
Analysis of survival data under the proportional hazards model.
Int.Stat.Rev., Vol. 43, 45-58.

2.- Breslow,N.E, y Crowley,J. (1974)
A large sample study of the table and product limit estimates under censorship.
Ann.Stat., Vol. 2, 437-453.

3.- Cox,D.R. (1972)
Regression models with life tables.
J.R.S.S., Ser. B, Vol. 34, 187-200.

4.- Hall,G.J., Rogers,W. y Pregibon,D. (1982)
Outliers matters in survival analysis.
The rand paper series, no. P-6761.

5.- Kalbfleisch,J.D. y Prentice,R.L. (1980)
The statistical analysis of failure time data.
N.Y., J.Wiley and Sons.

6.- Kaplan,E.L. y Meier,P. (1958)
Nonparametric estimation from incomplete observations.
JASA, Vol. 53, 457-481.

7.- Trejo V.,G.M.B. (1985)
La función de influencia en el análisis de datos de supervivencia.
Tesis de Maestría, IIMAS-UNAM.

CONSIDERACIONES DE TIPO METODOLOGICO EN RELACION AL ANALISIS DE UN ESTUDIO LONGITUDINAL DE CRECIMIENTO DE NINOS.

L.V.Schlaepfer, Ph.D.

Esta presentación consiste en una descripción parcial del análisis del crecimiento de dos series de niños seguidos longitudinalmente por el Dr. A. Chavez y su equipo (1), desde el nacimiento hasta los diez años de edad, a partir de 1968, en una comunidad rural del estado de Puebla. Los niños del grupo intervenido (n=14) recibieron suplementación alimentaria ad libitum durante todo el periodo de su crecimiento. Los ninos que forman parte del grupo control (n=16) subsistieron conforme a patrones dietéticos locales y deficientes. A continuación se pretende hacer hincapié en los problemas de tipo metodológico encontrados a lo largo del análisis y describir los resultados, en forma somera, solo en el caso de que ilustren algun aspecto de tipo interpretativo.

El crecimiento humano esta influenciado por una serie de factores en gran parte interdependientes, como son los de tipo genético, congénito, familiares, socio-económicos, así como de salud y de nutrición. El objetivo principal del estudio fué cuantificar la importancia relativa que tienen los factores nutrición, salud, sexo

(factor genético), y características al nacer (peso y longitud, representando a factores congénitos) sobre el crecimiento en peso y talla, de los niños de la muestra. Se utilizó el siguiente modelo de regresión multiple para relacionar el crecimiento con los factores mencionados:

$$\text{CRECIMIENTO} = c_0 + c_1 \text{ Nutrición} + c_2 \text{ Salud} + c_3 \text{ Sexo} + c_4 \text{ Características al nacer} + \text{error}$$

Crecimiento : Se describieron los datos de crecimiento de cada niño, por medio de una función matematica, cuyos parametros representarían sus características de crecimiento y resumirían la información contenida en la serie de sus mediciones (en promedio 60 mediciones por niño). La literatura ofrece dos modelos que parecieron particularmente adecuados para la descripción del crecimiento en el periodo de edad considerado: el de Count (linear, con un término logaritmico) y el de Jenss-Bayley (no linear, con un término exponencial). Se ajustaron los dos modelos a los niños del estudio y se comparó su rendimiento, en cuanto al grado de cumplimiento de los supuestos iniciales sobre la distribución de los errores y la homogeneidad de variancias, y a la bondad de ajuste.

El modelo de Jenss se mostró superior al de Count en todo respecto, tanto en el caso de la talla como en el del peso. Los ajustes observados, aunque buenos, fueron

inferiores a los de otro estudio llevado a cabo en niños sanos de Estados Unidos (2). Esto se puede deber a diferentes razones, incluso a una menor precisión de las mediciones, pero es más probable que se deba a los modelos mismos. En efecto, estos modelos fueron desarrollados para describir crecimiento "normal" y, en el estudio mencionado, fueron aplicados en niños sanos, que crecían conforme a las normas nacionales de E.U. En el presente caso, los niños presentan un patrón de crecimiento distinto, y es posible que el crecimiento de niños crónicamente desnutridos no sea normal, en el sentido de que no pueden ser descritos adecuadamente por los modelos planteados. Existen, tambien, auto-correlaciones entre los errores cuando las mediciones se efectuaron a un intervalo menor de 3 meses (o sea a partir de los 5 años de edad), lo que sugiere que un modelo auto-regresivo de primer orden pudiera ser más adecuado.

Sin embargo, ambos modelos explican por arriba del 95% de la variación en los datos lo que justifica la utilización de sus parámetros para resumir, eficazmente, la información contenida en la serie de observaciones. A continuación se presenta el modelo de Jenss, que fué el seleccionado para los análisis ulteriores:

$$y = a_0 + a_1 t - e^{(a2 + a3 t)} + e$$

donde: y = talla o peso observado

t = edad

e = error aleatorio.

El cuadro 1 muestra los valores promedio de los coeficientes del modelo de Jenss para el peso y la talla, y en el cuadro 2 se comparan los resultados de los análisis de variancia uni y multivariados de los mismos, por grupo y sexo. Se puede apreciar que el patrón de crecimiento es significativamente diferente entre los grupos intervenido y control pero, no entre sexos. Para el caso del peso, la diferencia se debe a que el grupo control exhibe valores mas bajos en los coeficientes a_0 y a_3. Los valores de los otros dos coeficientes son similares para los dos grupos. Esto sugiere que los niños control tienen inicialmente curvas más empinadas, y alcanzan el periodo de crecimiento linear a una edad más temprana (a_3 inferior) y, luego, crecen a la par de los niños suplementados (a_1 igual), aunque a un nivel más bajo. Para el caso de la talla, el grupo control exhibe, una vez más, valores inferiores a los del grupo intervenido para los parámetros a_0 y a_3 pero, aqui, su valor promedio de a_1 es algo superior, lo que pudiera significar un esfuerzo para compensar por la relativa brevedad de su fase de crecimiento rápido, mediante una velocidad de crecimiento lineal más alta. Al llegar a la entrada de la adolescencia, sin embargo, no han logrado aun compensar su retraso, puesto que

la diferencia en los promedios grupales de talla es, a los diez años, todavia de 5 cm.

Nutrición: Se comparó, en una primera instancia, el estado nutricional de los niños de los dos grupos mediante 1) indicadores de talla para la edad y peso para la talla, y 2) la presencia de evidencias de desnutricion clinica. Comparados con los niños del grupo control, los niños suplementados mostraron, a cada edad, y prácticamente sin excepciones, un estado nutricional superior, indicando la probable importancia de este factor sobre el crecimiento. Debido a la falta de datos de ingesta, se decidió utilizar la pertenencia a los grupos experimentales como indicador de nutrición (variable GRUPO). Esta es pues una variable categórica de dos niveles: el mejor y el peor nutrido.

Salud: La morbilidad se midió a travez de visitas familiares semanales, en las cuales se preguntaba el tipo de enfermedad o signo ocurrido y el numero de días con severidad de grado 1, 2, y 3(definidos según criterios de bienestar previamente establecidos).Se construyó, pero no se validó, el indice DURGRAV para considerar simultaneamente la duración y la gravedad de un episodio.

DURGRAV = (no. de días con severidad de grado 1 x 1) + (no. de días con severidad de grado 2 x 2) + (no. de días con severidad de grado 3 x 3)

El número de indices de morbilidad que se pueden derivar de la información existente es abrumador. Estos incluyen la frecuencia, duración y severidad de cada signo o enfermedad o de combinaciones de estos, en diferentes edades o grupos de edad, número de dias sanos, número y tipo de complicaciones, etc... Se optó por el siguiente método para seleccionar las variables adecuadas. Se compararon los promedios de todas las posibles variables, entre los dos grupos, por analisis de t-Student. Si no se encontró una diferencia significativa para una variable determinada, se procedió a su eliminación bajo el supuesto de que, en tal caso, no contribuiría a diferencias percibidas en el crecimiento de los dos grupos. Las variables que si fueron significativamente diferentes se mantuvieron. El grupo intervenido exhibió un mejor estado de salud de acuerdo a la mayoría de los indicadores, globalmente y en cada grupo de edad considerado. El análisis por grupos de edad reflejó la experiencia general indicada por la literatura de los procesos infecciosos en los paises en vías de desarrollo, según la cuál la incidencia de los mismos es baja durante el primer semestre de vida, aumenta dramáticamente en el segundo semestre y segundo y tercer años, y luego decrece paulatínamente durante el resto de los años prescolares, hasta alcanzar nuevamente valores bajos durante los años escolares. De esta manera, se decidió restringir las variables a los grupos de edad 6-12 meses y 1-5 años. El

cuadro 3 muestra las variables que quedaron seleccionadas en base a las consideraciones anteriores, con su codificación, y el cuadro 4 los promedios y desviaciones estándar de las mismas.

Como el numero de indicadores obtenidos parecía todavia muy grande, para los propósitos del analisis de regresión subsecuente, sobretodo en relación al tamaño de muestra, se ejecutó un análisis de componentes principales (ACP) para efectos de reducción de datos. El paquete estadístico SPSS que se utilizó para este fin, extrae automáticamente solo aquellas componentes cuyo valor singular es mayor o igual a 1, en este caso 3 componentes. La primera (CP1) mostró, como se esperaba, altas correlaciones con todas las variables. CP2 tuvo mayor correlación con la incidencia de diarreas en el periodo 6-12 meses de edad y con el número de días enfermos en el mismo periodo. CP3 se correlacionó principalmente con la incidencia de diarreas en el periodo 1-5 años de edad. Las componentes no se prestan para mayor esfuerzo interpretativo, pero sirven su propósito de reducción de datos, explican un 73.6% de la variación de los datos de morbilidad, y los valores que asumen para cada individuo se consideraron como sus características de salud.

Analisis de Regresion: Se ejecutó un análisis de regresión para cada coeficiente de peso y talla (modelo de Jenss) y ya sea para las 10 variables de salud originales o

para las componentes principales. La matríz de correlación de las variables independientes, llamó la atención hacia el hecho de que la variable GRUPO (i.e. nutrición) está significativa e inversamente relacionada con la mayoría de las variables de morbilidad (valores de r de Pearson entre -.52 y -.86). Esto dificultará la interpretación de los resultados de la regresión, como se verá más tarde. Las otras correlaciones fueron las esperadas, en función de la experiencia encontrada en la literatura y de los resultados de los análisis anteriores.

El cuadro 5 muestra los resultados de la regresión para los coeficientes de peso. Para a_3, GRUPO explica el 60% de la variación de los datos. La variable nutricional es el principal estimador para todos los coeficientes con excepción de a_1, confirmando así los análisis de variancia anteriores. Para a_2, la duración y gravedad de diarreas de 1 a 5 años, explica un 11% de la variación, dentro de cada grupo. El signo del coeficiente de esta variable es positivo, lo cual indica que, entre más grandes son la duración y gravedad de las diarreas, más bajo es el intercepto de la curva ponderal con el eje vertical o, vice-versa, entre más bajo sea el peso estimado al nacer, más grandes serán la duración y gravedad de las diarreas entre 1 y 5 años de edad, dentro de un mismo grupo.

Los mejores estimadores de a_1 son la incidencia de diarreas de 6 a 12 meses de edad, y la de infecciones respiratorias entre 1 y 5 años de edad. Pero aquí, si la incidencia de infecciones respiratorias está relacionada negativa y logicamente con la velocidad de crecimiento lineal, la incidencia de diarreas parece aumentar, en forma incongruente, paralélamente al incremento de esta. Es posible que esta variable haya entrado a la ecuación en forma fortuita, debido a las altas correlaciones que existen entre los indicadores de morbilidad, y al alto número de variables independientes en relacion al número de observaciones.

El cuadro 6 muestra los resultados correspondientes para talla. Una vez más, se puede notar la importancia de GRUPO como estimador para a_0 y a_2 (ya no a_3). Quizas haya llegado el momento de reflexionar sobre el hecho de que GRUPO es un indicador artificial del estado nutricional. Artificial, porque indica sencillamente que un niño pertenece a uno de dos grupos definidos al inicio del estudio. Estos grupos fueron pareados de tal manera que se diferenciaban únicamente por su ingesta alimentaria. Sin embargo, a medida que pasaba el tiempo, la diferencia nutricional entre los dos grupos resultó en otras diferencias, que no fueron posibles de controlar experimentálmente. Asi, los niños mejor nutridos se

volvieron niños más sanos que los peor nutridos, y esto esta confirmado por las altas correlaciones encontradas entre GRUPO y los indicadores de morbilidad. GRUPO es pués una variable que refleja, al mismo tiempo, estado nutricional y estado de salud. Quizá variables de ingesta alimentaria hubieran permitido una mejor discriminación entre los efectos de la salud y de la nutrición sobre el crecimiento.

Otra contribuyente importante a la variación en crecimiento en estatura, es la incidencia de diarreas entre 1 y 5 años de edad. Si se recuerda que el patrón de crecimiento en talla de los niños del grupo control está asociado con valores más pequeños para a_3 y mas grandes para a_1, comparados con los del grupo intervenido, se verá que altas incidencias de diarrea durante el periodo de 1 a 5 años de edad, estan relacionadas con este patrón deficiente de crecimiento.

El factor sexo se vuelve importante en la determinación de la variación en el crecimiento en talla. Dentro de un mismo nivel de incidencia de diarreas (1-5 anos), ser mujer parece implicar valores más bajos de la talla estimada al nacer, periodos más breves de crecimiento rápido inicial y velocidades lineales más pequeñas, en comparación con el sexo masculino. Esta relación entre sexo y crecimiento en estatura no es aparente a partir de la correlación univariada entre ambos. Sin embargo, la

correlación parcial entre a_1 y sexo se incrementa de .208 a .432, cuando se ajusta por la incidencia de diarreas de 1 a 5 años de edad. En igual forma se magnifican las correlaciones entre sexo y a_2 y a_3, respectivamente. Los análisis de variancia tampoco mostraron un factor sexo significativo.

Los cuadros 7 y 8 muestran los resultados de la regresiones, en las cuales las 3 CP remplazaron a las 10 variables crudas de morbilidad. No aparecen cambios importantes en relacion a los resultados anteriores. Ninguna variable nueva aparece en las ecuaciones. La diferencia más importante es la desaparición de la varible sexo de entre los predictores de los coeficientes a_2 y a_3. Este resultado es más acorde a los análisis de variancia.

Este análisis arrojo algunas indicaciones sobre las posibles causas de las deficiencias de crecimiento encontradas en la muestra de niños, sobre todo los peor nutridos. Sin embargo, no es conclusivo. La razón primordial es probablemente la interrelación entre los coeficientes del modelo. La utilización de alguna metodología capaz de considerar la relación de las variables independientes con el conjunto de los coeficientes, en vez de con cada uno en particular, sería de gran utilidad.

Cuadro 1

Estimadores de los parámetros del modelo de Jenss ajustado a la talla y al peso, respectivamente, para el grupo intervenido (n =14) y el grupo control (n=16). Tezonteopan, Pue

Parámetros	Media	Error Estándar	Mínimo	Mediana	Máximo
Grupo Experimental					
a_0	93.372	2.467	79.591	91.842	116.003
$a_1 \times 1000$	9.606	.544	5.470	9.445	13.000
a_2	3.721	.053	3.364	3.697	4.127
$a_3 \times 1000$	-1.658	.127	-2.680	-1.710	-0.910
Grupo Control					
a_0	76.769	2.139	69.027	73.430	95.251
$a_1 \times 1000$	12.774	.688	6.830	12.795	16.140
a_2	3.219	.062	2.956	3.131	3.737
$a_3 \times 1000$	-4.455	.305	-5.430	-2.260	-0.800
PESO					
Parámetros	Media	Error Estandar	Mínimo	Mediana	Máximo
Grupo Experimental					
a_0	10.307	1.609	5.460	8.755	28.896
$a_1 \times 1000$	5.106	.632	-.390	5.270	10.320
a_2	1.753	.149	0.902	1.673	3.189
$a_3 \times 1000$	-5.460	1.385	-19.860	-4.300	-0.360
Grupo Control					
a_0	5.557	.175	4.731	5.329	6.850
$a_1 \times 1000$	5.109	.179	3.680	5.050	6.740
a_2	1.040	.061	0.633	1.010	1.549
$a_3 \times 1000$	-17.550	1.217	-25.780	-18.300	-6.190

Cuadro 2

Análisis de variancia uni y mutivariado de los coeficientes del modelo de Jenss ajustado al peso y a la talla, respectivamente. Tezonteopan, Pue.

	Univariado			Multivariado		
	Factores	F	p	Factores	Approx.F	p
Peso						
a_0	Grupo	9.8	<.0041			
	Sexo	.7	<.4130			
a_1	Grupo	.0	<.9962			
	Sexo	.9	<.3459	Grupo	36.2	<.0001
				Sexo	1.3	<.3099
a_2	Grupo	22.8	<.0001			
	Sexo	2.7	<.1131			
a_3	Grupo	42.2	<.0001			
	Sexo	.2	<.6347			
Talla						
a_0	Grupo	26.2	<.0001			
	Sexo	.8	<.2870			
a_1	Grupo	13.3	<.0011			
	Sexo	2.7	<.1155	Grupo	17.0	
				Sexo	1.4	
a_2	Grupo	37.8	<.0001			
	Sexo	1.8	<.1877			
a_3	Grupo	5.7	<.0241			
	Sexo	3.4	<.0758			

Cuadro 3

Lista de indicadores de morbilidad seleccionados para el análisis de regresión con su codificación. Tezonteopan, Pue.

Diarrea	
Incidencia en el periodo de edad 6-12 meses	DIN612
Incidencia en el periodo de edad 1-5 años	DIN15
Indice DURGAV entre 6 y 12 meses de edad	DD612
Indice DURGRAV entre 1 y 5 años de edad	DD615
Infecciones respiratorias	
Incidencia en el periodo de edad 6-12 meses	RIN612
Incidencia en el periodo de edad 1-5 años	RIN15
Indice DURGRAV entre 6 y 12 meses de edad	RD612
Indice Durgrav entre 1 y 5 años de edad	RD615
No. de días enfermos entre 6 y 12 meses de edad	ID612
No. de días enfermos entre 1 y 5 años de edad	ID15

Cuadro 4

Promedios y desviaciones estándar de los indicadores de morbilidad seleccionados, por grupo. Tezonteopan, Pue.

Variables	Grupo Control	Grupo Intervenido
6-12 meses de edad		
Diarrea		
-No. ataques	5.1 (2.1)	4.3 (1.7)
-DURGRAV	13.0 (3.6)	8.7 (2.0)
Infecciones respiratorias		
-No. ataques	3.2 (1.2)	1.9 (0.8)
-DURGRAV	14.5 (5.6)	11.2 (5.1)
No. dias enfermos	85.1 (16.6)	54.6 (12.3)
1-5 años de edad		
Diarrea		
-No. ataques	28.7 (6.6)	18.2 (4.3)
-DURGRAV	14.6 (2.1)	9.5 (2.3)
Infecciones respiratorias		
-No. ataques	18.5 (4.5)	12.3 (3.1)
-DURGRAV	16.0 (1.7)	12.5 (2.2)
No. dias enfermos	623.8 (102.5)	344.4 (59.7)

Cuadro 5

Análisis de regresión múltiple para los estimadores de los parámetros de la curva de peso (modelo de Jenss) casos = 30. Tezonteopan, Pue.

Variable dep.	Variables indep. por orden de entrada	Coeficiente de regr. b	Error Estandar de b	Valor de t	Signif. de t	R^2
a_0	GRUPO	4.750	1.513	-3.1	p<.005	.260
	(Constante)	10.307	1.105	9.3	p<.0001	
a_1	DIN612	.540	.124	4.4	p<.0005	.298
	RIN15	-.138	.048	-2.9	p<.01	.164
	(Constante)	4.706		5.3	p<.0001	
a_2	GRUPO	1.128	.216	-5.2	p<.0001	.435
	DDG15	.082	.033	2.5	p<.05	.108
	(Constante)	.969	.327	3.0	p<.01	
a_3	GRUPO	12.090	1.835	-6.6	p<.0001	.608
	(Constante)	-5.460	1340	-4.1	p<.0005	

Cuadro 6

Análisis de regresión múltiple para los estimadores de los parámetros de la curva de talla (Modelo de Jenss)
casos = 30. Tezonteopan, Pue.

Variable dep.	Variables indep. por orden de entrada	Coeficiente de regr. b	Error Estándar de b	valor de t	Signif. de t	R^2
a_0	GRUPO	16.603	3.248	-5.1	p<.0001	.483
	(Constante)	93.372	2.372	39.4	p<.0001	
a_1	DIN15	.250	.054	4.6	p<.0005	.341
	SEXO	2.049	.824	2.5	p<.05	.123
	(Constante)	4.249	1.512	2.8	p<.01	
a_2	GRUPO	.319	.103	-3.1	p<.005	.567
	DIN15	-.019	.007	-2.7	p<.05	.061
	SEXO	-.161	.076	-2.1	p<.05	.055
	(Constante)	4.152	.153	27.14	p<.0001	
a_3	DIN15	-.072	.021	-3.4	p<.005	.204
	SEXO	-.805	.318	-2.5	p<.05	.153
	(Constante)	.054	.583	.9	n.s.	

Cuadro 7

Análisis de regresión múltiple para los estimadores de los parámetros de la curva de peso (modelo de Jenss) con las componentes principales remplazando los indicadores crudos de morbilidad. Tezonteopan, Pue.

Variable dep.	Variables indep. por orden de entrada	Coeficiente de regr. b	Error Estándar de b	Valor de t	Signif. de t	R^2
a_0	GRUPO	-4.750	1.513	-3.1	.0040	.260
	(Constante)	10.307	1.105	9.3	.0000	
a_1	CP3	.842	.271	3.1	.0044	.219
	CP2	-.676	.266	-2.5	.0169	.151
	(Constante)	5.108	.250	20.4	.0000	
a_2	GRUPO	-6.931	.153	-6.1	.0000	.435
	CP2	.246	.081	3.0	.0052	.144
	(Constante)	1.869	.106	3.0	.0063	
a_3	GRUPO	-14.403	1.887	-7.6	.0000	.808
	CP2	2.624	1.000	2.6	.0141	.660
	(Constante)	-4.226	1.306	-3.2	.0032	

Cuadro 8

Análisis de regresión múltiple para los estimadores de los parámetros dela curva de talla (modelo de Jenss) con las 3 componentes principales remplazando los indicadores crudos de morbilidad. Tezonteopan, Pue.

Variable dep.	Variables indep. por orden de entrada	Coeficiente de regr. b	Error Estándar de b	Valor de t	Signif. de t	R^2
a_0	GRUPO	-19.946	3.478	-5.7	.0000	.483
	CP2	3.793	1.842	2.1	.0493	.070
	(Constante)	95.155	2.407	39.5	.0000	
a_1	GRUPO	4.683	.825	5.7	.0000	.309
	CP2	-1.575	.438	-3.6	.0013	.176
	SEXO	1.770	.733	2.4	.0230	.094
	(Constante)	7.854	.721	10.8	.0000	
a_2	GRUPO	-.597	.087	-6.9	.0000	.567
	CP2	.108	.046	2.3	.0275	.073
	(Constante)	4.152	.153	27.14	.0000	
a_3	GRUPO	-.797	.348	-2.3	.0296	.159
	(Constante)	-1.658	.254	-6.5	.0000	

UNA METODOLOGIA PARA CLASIFICACION DE SUELOS

Rosa María Ochoa A.
IIMAS-UNAM.

RESUMEN

Cuando se lleva a cabo una clasificación de suelos sin tomar en cuenta la posición geográfica de los puntos muestrales, puede dar como resultado mapas de suelos muy fragmentados y con fronteras intrincadas. Para resolver este problema, se introducirá el uso de una función exponencial para alisar datos que depende de la distancia geográfica de los puntos muestrales.

Se presentará el uso de Componentes Principales, Coordenadas Principales, el Semivariograma y algunos métodos jerárquicos aglomerativos como herramientas para producir mapas alisados. Y por último se dará un criterio objetivo que garantiza consistencia con las observaciones determinando el grado de alisamiento. La metodología propuesta se aplica satisfactoriamente a datos provenientes de campos experimentales de la Universidad de Chapingo produciendo asi mapas de clasificación.

1 INTRODUCCION

Un problema común en casi todos los levantamientos de suelos para propósitos generales, es el de obtener un mapa de áreas contiguas pero disjuntas,de tal manera que dentro de cada una de las áreas, el suelo sea de la misma clase con respecto al perfil del suelo y en parcelas de tierra razonablemente grandes. Por lo general resultan varias áreas separadas de cada tipo de clase.

Cuando sólamente se esta interesado en una propiedad del suelo en particular, se pueden crear clases de suelo dividiendo su rango de variación a ciertos puntos fijos. A este proceso se le conoce como "disección", y la única consideración es que la clasificación sea util para el propósito que se tiene en mente. Pero cuando se incrementa el número de propiedades de interés, entonces nos enfrentamos al problema de clasificarlo tomándolas en cuenta a todas ellas a la vez, y llevando a cabo una disección simultánea de cada propiedad nos daría como resultado demasiados grupos. El uso de computadoras ha estimulado a investigar la solución a este problema y ahora existen varios métodos numéricos para clasificación de datos de tal manera que no se crean

muchos grupos y que reflejan la relación que existe entre los datos, siendo generalmente muy útil.

Dentro de los métodos jerárquicos para clasificación, los métodos jerárquicos aglomerativos son los mas populares para clasificar suelos debido a que en base a su estructura son los que se han acomodado mejor a clasificar este tipo de datos. Por lo que son los que se van a usar en este trabajo para producir mapas. Pero hasta aquí, el problema de clasificación aún no se ha resuelto. Cuando se hacen descripciones del suelo como parte de un levantamiento de un área, existe una característica adicional de los datos, la cual se llama "localización geográfica" de cada punto muestral. La creación de clases de perfiles de suelos, sin tomar en cuenta el arreglo espacial, puede causar una fragmentación del área en pequeñas parcelas separadas geográficamente o en parcelas más grandes que poseen fronteras muy intrincadas y consecuentemente, la clasificación no tiene un valor práctico. El levantador de suelos trata de crear clases dentro de las cuales las variaciones de las características del suelo sean lo suficientemente pequeñas para el propósito deseado, y que resulten parcelas individuales compactas que sean razonablemente grandes. Si es posible, él evitará una clasificación que divida el área en parcelas más pequeñas o con fronteras intrincadas, más de lo necesario.

Webster y Burrough(1972a) introdujeron un proceso llamado "alisamiento" por medio del cual los problemas de fragmentación y fronteras intrincadas pueden ser reducidos sin alterar seriamente la información original.

2 ALISAMIENTO

Mientras más grandes sean las áreas de suelo que pueden ser tratadas de la misma manera, y mientras más suaves sean sus fronteras, más libertad se tendrá para planear lo que se quiera hacer con dichas áreas facilitando su manejo. Para lograr esta situación, sería deseable que no existieran fragmentación ni fronteras intrincadas.

Los mapas de suelos se basan generalmente en observaciones multivariadas de la forma $X_i = (x_1, x_2, \ldots, x_p)$ donde las x_i son variables aleatorias discretas o continuas, que fueron tomadas posiblemente a diferentes profundidades en el suelo y en los vértices de una red regular, donde se hace la suposición de que el punto es representativo del cuadrado en el que cae y que ha sido medido sin error.

En la práctica, la mayoría de las propiedades del suelo, muestran una variación espacial a escalas muy diferentes. Puede haber tendencias a gran escala sobre distancias de cientos de metros, o comparativamente fronteras muy marcadas entre formaciones geográficas diferentes. Además, aún dentro de tales formaciones, puede haber variación local a un rango muy corto, aún entre muestras tomadas a muy pocos centímetros de separación. Este tipo de variación es reflejada en lo que commmente llaman efecto "nugget" o efecto pepita, que se verá más adelante. Entonces, picos locales o valles en el mapa pueden corresponder a efectos de la variación a escalas muy pequeñas exagerando así la irregularidad de la propiedad representada. Es este tipo de efecto, el que indica la necesidad de alisar, usualmente se presenta en forma exagerada en mapas de clasificación de suelos.

La técnica de alisamiento que se necesita para alisar fronteras irregulares y evitar fragmentación, no debe implicar que puntos que se encuentran bastante separados en el espacio sean observados como disímiles. Ya que suelos similares pueden pertenecer a la misma clase, independientemente de que se encuentren en áreas diferentes, separadas por suelos de un tipo muy distinto. Además, uno debe tener cuidado de que al ajustar los valores observados, o al reclasificar suelos cuando éstos se encuentran en puntos aislados o en áreas muy pequeñas, no se descarte información genuina contenida en las observaciones. Cuando los datos indican que un área muy pequeña, aun representada por una sola observación difiere muy marcadamente de sus vecinos geográficos, este hecho, no debe ser suprimido por un sobrealisamiento. De aquí surge entonces, el problema final de encontrar un criterio objetivo para detener el procedimiento de alisamiento y así producir un mapa consistente con las observaciones, dando lugar a fronteras regulares entre regiones diferentes, como se verá mas adelante.

La técnica de alisamiento utilizada en este trabajo, utiliza la idea de medidas de disimilaridad, que es usada comunmente en edafología. Muchas de estas medidas han sido propuestas; entre otras se encuentran: el coeficiente de similaridad de Gower (Gower, 1971), la "Canberra Metric" (Lance y Williams, 1967) y algunas medidas de distancia. Una buena recolección de estas medidas es dada en Sneath y Sokal (1973).

Frecuentemente la clasificación de suelos se lleva a cabo utilizando análisis de conglomerados basándose en medidas de disimilaridad. Para evitar irregularidades en el mapa, se vuelve necesario ajustar la medida de disimilaridad de acuerdo a la distancia geográfica en el suelo, de tal manera que, la disimilaridad entre observaciones que están muy cerca se reduce o se incrementa para observaciones que estan lejos. El objetivo es reducir disimilaridades muy improbables entre puntos vecinos geográficamente; no se desea por supuesto incrementar la disimilaridad entre observaciones que están

lejos, de tal manera que suelos similares en diferentes distritos no puedan ser clasificados en la misma categoría. Se requiere una función de distancia que cambie rápidamente para distancias cortas, pero que sea aproximadamente constante para distancias grandes. Webster y Burrough (1972b) propusieron una función muy adecuada, la cual alisa irregularidades locales y elimina pequeñas parcelas fragmentadas, mientras que al mismo tiempo, reconoce parcelas que están muy separadas y que no son similares para colocarlas en la misma clase de suelo. Esto se hace variando la contribución que aporta el índice de disimilaridad por medio de la localización, de acuerdo a la distancia geográfica entre pares de sitios. Y así el análisis de conglomerados se va a llevar a cabo basándose en las disimilaridades transformadas, que denotaremos por d^*_{ij} y definida como:

$$d^*_{ij} = d_{ij}\left(1 - e^{-\frac{\Delta_{ij}}{w}}\right)$$

donde:

d_{ij} es una medida de disimilaridad entre los puntos i y j; la cual en nuestro caso, es la distancia euclidiana entre los puntos i y j basada en las medidas originales estandarizadas tomadas en cada punto muestral.

Δ_{ij} es la distancia geográfica entre los puntos i y j.

w es un peso asignado a la distancia geográfica Δ_{ij}.

Debe uno hacer notar, que aun si las dimilaridades originales pueden ser representadas como distancias euclidianas en algún espacio de dimensiones conveniente, esto no será cierto en general de las disimilaridades ajustadas.

La idea general para evitar fragmentación es entonces, modificar la matriz de disimilaridades, tomando en cuenta la localización de los puntos muestrales en el suelo y asi llevar a cabo una clasificación posterior. Hasta aquí, el parámetro w, el cual es el grado de alisamiento, esta indeterminado.

El problema con este tipo de métodos es el de determinar w evitando un sobrealisamiento. En la sección 5 se discutirá un criterio estadístico objetivo para determinar w.

3 REDUCCION DE DIMENSIONALIDAD

Varias técnicas matemáticas han sido propuestas con el fin de reducir las características más importantes de un conjunto de datos en un número menor de variables. El análisis de componentes principales es tal vez el más importante. Sin embargo, Gower(1966) muestra que existe otra técnica dual al análisis de componentes principales, llamada "análisis de coordenadas principales", que se basa en las medidas de disimilaridad en lugar de los valores de las variables.

También demostró que las coordenadas principales coinciden con las componentes principales (la primera coordenada con el primer componente y as'i sucesivamente) cuando en el análisis de coordenadas principales se usa cono matriz de disimilaridad la matriz de distancias euclidianas, y el análisis de componentes principales se lleva a cabo con la matriz de correlación de las variables originales.

El análisis de coordenadas principales ha sido ampliamente usado para definir estructuras en varias poblaciones de suelos, como se puede mencionar a Rayner(1966,1969), Campbell et al.(1970) y a Webster y Butler(1976).

Holland(1969) ha utilizado el análisis de componentes principales para clasificar sus puntos muestrales en 3 clases: alta, mediana y baja dependiendo de los valores que toma el primer componente principal y ha mostrado sus resultados como mapa de clasificación.

En este trabajo, se llevará a cabo el análisis de componentes principales con la matriz de correlación de los datos originales; análisis de coordenadas principales utilizando las matrices que resulten de ajustar las disimilaridades y siempre y cuando el primer componente principal tanto como la primera coordenada principal para cada uno de los análisis que se hagan sobre las matrices de disimilaridades ajustadas, representen la mayor parte de la variabilidad presente total contenida en los datos, serán variables muy razonables por considerar. Por lo que con ellas se van a producir mapas de clasificación como los hizo Holland(1969) y también utilizando métodos jerárquicos aglomerativos.

4 EL SEMIVARIOGRAMA.

En la práctica, las clases de suelos se reconocen a partir de la combinación del perfil del suelo y distribución espacial. Cuando los datos de puntos vecinos suelen ser similares, mientras que los datos de puntos distantes difieren mas seguido, entonces se dice que los datos son espacialmente dependientes.

Aunque este problema se resuelve usualmente de una manera intuitiva mas que analítica, avances recientes en geoestadística han dado los medios para medir esa dependencia espacial y sus consecuencias para clasificación. De tal manera, la teoría de variables regionalizadas construye una función complementaria conocida como "semivariograma), en la cual, la semivarianza es una medida de dependencia espacial, i.e. es una medida de la disimilaridad en promedio, entre puntos a una distancia dada. El semivariograma estimado en la dirección para la distancia h esta expresado por:

$$\gamma^{*}(h,\alpha) = \frac{1}{2}\, n(h,\alpha) \sum_{i=1}^{n(h,\alpha)} \{x(i) - x(i+h)\}^2$$

donde h - es la distancia que separa a dos pares de puntos x(i) y x(i+h).

n(h, α) - son los pares de puntos separados por el vector h.

Hay varios modelos de semivariogramas teóricos, (Journel y Huijbregts 1978), tal vez el mas simple es el modelo lineal que

se muestra en la figura 4.1. El "verdadero" semivariograma, por definición h empieza en 0 puesto que es imposible tomar dos muestras más cercanas que a distancia 0. Considérese el caso en que h=0. Se toman dos muestras exactamente en el mismo lugar y se miden sus valores. La diferencia entre los dos debe de ser 0, por lo que el semivariograma debe de pasar siempre a través del origen de la gráfica.

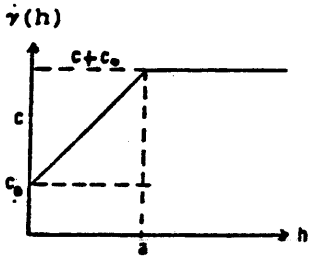


Fig. 4.1 Modelo Lineal de Semivariograma

En la mayoría de los casos esto no sucede, la regularidad en el espacio de la variable de interés, se relaciona estrechamente con el comportamiento del semivariograma muy cerca del origen.Por lo general el semivariograma no tiende a cero al tender h a cero, sino que aparece un valor positivo C_0 que intersecta al eje de las ordenadas como se muestra en la Fig. 4.1. A este tipo de efecto se le conoce como efecto "nugget" o efecto "pepita" y a C_0 se le conoce como varianza "nugget" o varianza "pepita", y representa la variación presente dentro de distancias menores que el intervalo muestral, i.e. es una medida de variación a muy corta escala.

Campbell (1978) usó el semivariograma por primera vez en estudios de suelos, recientemente, Burgess y Webster (1980a,b) y Hajrasuliha et. al. (1980) lo han aplicado para la interpolación y estimación de mapas de suelos y Webster(1981) lo utilizó para determinar la dependencia espacial y clasificar el suelo sobre un transecto. Matheron (1965,1971)y Journel y Huijbregts(1978) usan semivariogramas principalmente para caracterizar estructuras de variables dentro de depósitos.

En este trabajo, se va a utilizar el semivariograma como herramienta para estimar mapas de suelos utilizando el primer componente principal basado en las observaciones originales y la primera coordenada principal producida por las matrices de disimilaridades ajustadas, i.e. las coordenadas principales alisadas.

5 CRITERIO PARA DETERMINAR EL GRADO DE ALISAMIENTO.

En todos los métodos de alisamiento el problema principal es saber cuándo detener el procedimiento. Los datos no alisados muestran irregularidades asociadas a variaciones de escala muy pequeña y a errores de medición. Estas irregularidades causan fragmentación y fronteras complejas en los mapas de suelos, debido a que los subconjuntos del espacio de caracteres asignados a diferentes clases de suelo tienen límites muy abruptos, o bien muy delineados, y el suelo cerca de esos límites puede cambiar de tipo como resultado de variaciones muy pequeñas. Por otra parte, un sobrealisamiento producirá un mapa muy regular a la vista, en el cual la información real se ha perdido. Por ejemplo, puede haber irregularidades genuinas en las fronteras y parches de suelo aislados reales de diferentes tipos; lo que no debe pasar inadvertido, es que el alisamiento no debe perder esa información.

El grado de alisamiento se puede escoger subjetivamente (Webster, 1977), de tal manera que el mapa resultante se vea un tanto regular, o parcialmente regular, pero que no pierda información real de acuerdo con los conocimientos del edafólogo. Así, distintos grados de alisamiento servirán de guía al edafólogo para hacer un mapa realista.

Se puede uno preguntar, si los datos proveen algún criterio objetivo acerca de la distinción entre variación local que se pueda omitir, y diferencias más grandes que estén asociadas con distinciones de importancia práctica. El peso "w" se deberá escoger de tal manera que los cambios en las disimilaridades entre individuos (con una corrección de escala según se requiera), sean compatibles con la variación local en las propiedades medidas.

Para una sola propiedad, la variación local se puede estimar mediante el semivariograma. Donde hay dependencia local fuerte, la varianza pepita si es que la hay, estima la variabilidad a una escala más pequeña que la distancia del intervalo muestral. Donde la dependencia local no es fuerte; es casi siempre posible obtener alguna idea del límite superior de la varianza pepita y éste probablemente será un poco menor que la varianza para las distancias más cortas. Ahora, el alisamiento causa cambios en el valor de la variable; y es fácil calcular la media del cuadrado de estos cambios a la que se llamará "varianza de los cambios o de las diferencias", puesto que la media en ambos casos es cero. Si la varianza de los cambios es mucho más grande que la varianza pepita, se puede decir con certeza que los ajustes del alisamiento han hecho cambios incompatibles con los datos observados en

las variaciones a escala muy pequeña y que el mapa resultante está probablemente sobrealisado.

El alisamiento exponencial sobre la matriz de disimilaridad discutido en la sección 2,no nos permite aplicar esta técnica a variables individuales. Las disimilaridades d_{ij} no definen valores ajustados de pH o contenido de barro, por dar un ejemplo. Pero si se lleva a cabo un análisis de componentes principales sobre las variables originales y análisis de coordenadas principales sobre las matrices de disimilaridades ajustadas, el primer componente y la primera coordenada principal son unas variables razonables por considerar, puesto que representan la mayor parte de la variabilidad total contenida en los datos. Es facil construir un semivariograma para cada una de estas variables. Ahora el alisamiento va a producir cambios en la primera coordenada principal, pero, siempre y cuando el primer valor propio sea considerablemente más grande que el segundo, esta relación se mantendrá para un grado de alisamiento bastante razonable.

Si como resultado del alisamiento, la varianza de los cambios en la primera coordenada principal es mayor que la varianza pepita del primer componente principal (i.e. datos no alisados), tendremos entonces una indicación de sobrealisamiento.

Esto da una guía objetiva para el parámetro de alisamiento apropiado "w". Notemos que este criterio no es un sustituto del concimiento del edafólogo, y no indica un valor parámetro "w" como el "el mejor". Sin embargo, sirve para comprobar.

Si para cierto grado de alisamiento se observan cambios incompatibles con la varianza pepita en la primera coordenada principal, el estadístico entonces puede por lo menos decir: "Esto está distorcionando la información que aporta los datos y ajustando mucho más que fluctuaciones menores asociadas con variación a muy corta escala".

Esto no tiene importancia al realizar un análisis de conglomerados y al construir mapas, pero sí exagerará las diferencias entre los valores de la coordenada principal alisada y la no alisada. Por lo que es mejor escalar los valores alisados para tener la misma variación que los no alisados; y así la varianza de los cambios entre los valores no alisados y los valores alisados escalados,. puede ser calculada fácilmente y luego comparada con la varianza pepita.

6 ANALISIS DE LOS DATOS.

La metodología propuesta se usó para producir mapas de clases de datos colectados en la Universidad de Chapingo. El área de estudio es rectangular y mide 160 m x 200 m. Las muestras fueron colectadas utilizando muestreo sistemático. El área se dividió en una red regular de 20 m el intervalo, dando 8 puntos muestrales de Norte a Sur y 10 puntos muestrales de Oeste a Este, lo cual hace un total de 80 puntos muestrales.

Las propiedades medidas en cada punto muestral fueron: contenido de arena y barro para la textura del suelo; porcentaje de humedad, pH y conductividad eléctrica.

Con estos datos se formó la matriz de distancias euclidianas entre las observaciones de estos puntos. Debido a que las mediciones obtenidas tenían escalas muy diferentes, fue necesario estandarizar las variables; por lo que la matriz de distancias euclidianas entre los puntos i y j se formó de las distancias denotadas por d_{ij} dadas por:

$$d_{ij} = \left\{ \sum_{k=1}^{p} (y_{ik} - y_{jk}) \right\}^{1/2} \qquad (6.1)$$

donde

$$y_{ik} = \frac{x_{ik}}{\sigma_k}$$

x_{ik} es la k-ésima observación original al punto i,

σ_k es la desviaci'on estándar de la k-ésima variable.

Se transformó a la matriz d_{ij} de acuerdo a la regla:

$$d^{*}_{ij} = d_{ij}\left\{1 - e^{-\frac{\Delta_{ij}}{w}}\right\}$$

para w = 0, 0.5, 1, 1.5, 2, 2.5, 3 y 10.

Se llevó a cabo un análisis de componentes principales sobre la matriz de correlación del conjunto de datos originales, lo cual es equivalente a realizar un análisis de coordenadas principales sobre la matriz de disimilaridades transformadas d^{*}_{ij} para w=0.

El primer componente principal representó un porcentaje de varianza muy alto por lo tanto es una buena variable por considerarse; por lo que se puede proseguir con el análisis.

Se practicaron análisis de coordenadas principales con las matrices d^{*}_{ij} para los diferentes valores de w. La primera coordenada principal dominaba al incrementarse el valor de w y el porcentaje de varianza para esta coordenada se mantuvo mucho más alto que el de las otras coordenadas principales. Por lo tanto, retendrá su identidad para un grado de alisamiento razonable. Entonces, se puede usar la primera coordenada principal para continuar con el análisis y producir mapas, al igual que se hace utilizando el primer componente principal.

Una vez que se han obtenido las matrices d^{*}_{ij}, se hace necesario encontrar el valor apropiado de w de acuerdo a la Sección 5 para calcular la varianza de los cambios entre los datos alisados y no alisados. Recordando que es necesario escalar los valores alisados para tener la misma escala de variación que los no alisados, y así hacerlos comparables. Por lo que, multiplicando cada componente de la primera coordenada principal por un "factor de ajuste", que se define en la fórmula 6.4, se obtienen los valores alisados escalados. Entonces, se procede a calcular la "varianza de los cambios ó diferencias" entre los datos no alisados y los datos alisados escalados para cada uno de los valores de w y a ser comparadas con la varianza pepita que se obtenga del semivariograma para el primer componente principal. La varianza de los cambios está definida por la fórmula siguiente:

$$Var_C = \frac{\sum_{i=1}^{n}\left\{Comp_I(i) - Coord_I(i)\ (Factor\ de\ Ajuste)\right\}^2}{n-1} \qquad (6.3)$$

donde

$Comp_I(i)$ denota el i-ésimo elemento del primer componente principal.

$Coord_I(i)$ denota la i-ésima coordenada de la primer coordenada principal.

El factor de ajuste para la coordenada I se define por:

$$\left\{ \text{Varianza } Comp_I \Big/ \text{Varianza } Coord_I \right\} \qquad (6.4)$$

por lo que, el criterio descrito en la Sección 5 es el siguiente:

Si la varianza de los cambios es $^a=$ varianza pepita , implica evidencia de sobrealisamiento.

Como se verá posteriormente en la sección 6.2 (Fig. 6.2) el estimador de la varianza pepita para el primer componente principal fué aproximadamente de 0.5. Para w=3, la varianza de los cambios fué igual a 0.35. Esto sugiere que 3 puede ser un grado de alisamiento bastante aceptable; mientras que para w=10, el valor de la varianza de las diferencias fué igual a 0.77. Siendo éste más grande que el valor de la varianza pepita (0.5), lo cual sugiere que este grado de alisamiento está distorsionando los datos y que han sido sobrealisados.

6.1 Mapas de clasificación: alto, mediano y bajo

Con el objeto de mostrar claramente el efecto que está produciendo el alisamiento exponencial, se hicieron mapas de clasificación de acuerdo a valores altos, medianos y bajos para los datos no alisados (i.e. w=0) y para los otros valores de w. En la Fig. 6.1 se muestran estos mapas para los datos no alisados, w=3 y w=10, en donde se puede apreciar que prevalecieron 3 clases (para valores altos, medianos y bajos). Los valores altos de acuerdo al primer componente principal están asociados con presencia de arcilla (mal drenage) y los valores bajos con presencia de arena (buen drenage). Además se puede observar

el efecto de que valores aislados muy negativos rodeados por valores medianos, se transformaron en menos negativos y algunos de ellos para w=10 hasta cambiaron de signo debido al sobrealisamiento. Lo mismo sucede para valores aislados muy positivos rodeados de negativos y valores medianos rodeados de valores ya sea muy negativos o muy positivos. Estos mapas representan la mayor parte de la variación presente en el área de estudio, de acuerdo a la distribución del primer componente principal para los datos no alisados y la primera coordenada principal escalada para los datos alisados.

6.2 Semivariogramas

El primer componente principal representó un porcentaje de la varianza total muy alto; por lo que se obtuvo con éste un semivariograma experimental para las cuatro direcciones principales: Norte-Sur (N-S), Este-Oeste (E-O), Noroeste-Sureste (NO-SE) y Noreste-Suroeste (NE-SO). Cualquier curva suave que se ajuste a este semivariograma dará una varianza pepita diferente de cero, que representa la variación dentro del intervalo muestral. El estimador de la varianza pepita obtenido por extrapolación fue de 0.5, o algo menor a 0.5, lo cual sugiere que puntos adyacentes no son muy similares.

De la misma manera se produjeron semivariogramas para w=3 y w=10 utilizando la primera coordenada principal escalada. En la Fig. 6.2 se pueden apreciar estos 3 semivariogramas, donde se observa claramente el efecto de alisamiento; al incrementarse w, la variabilidad para distancias cortas se reduce considerablemente y presenta un efecto pepita igual a cero, (ver para w=10). Por lo que se puede decir que como resultado del alisamiento los puntos adyacentes son muy similares.

6.3 Mapas de Clasificación : Técnicas de conglomerados.

Utilizando la estrategia de vecinos más lejanos, se llevaron a cabo análisis de conglomerados sobre los datos no alisados y los alisados para cada valor de w. En la Fig. 6.3 se pueden observar los mapas resultantes para los datos no alisados y para w=3 y w=10 con 5, 4 y 3 clases. En la parte superior se reprodujeron los mapas de la Fig. 6.1.

Al incrementarse el valor de w se hace evidente que hay menos fragmentación entre las clases. Se puede decir que existe una clase muy bien definida en la parte superior izquierda que puede corresponder a los valores muy positivos asociados con suelos arcillosos (drenage muy pobre) de la Fig. 6.1. Así como se puede asociar la clase que se encuentra en la parte superior derecha con algunos de los valores medianos que definen otra clase y finalmente, la clase que se observa en la parte inferior derecha puede ser asociada con los valores muy negativos que corresponden a suelos arenosos (buen drenage). Las otras

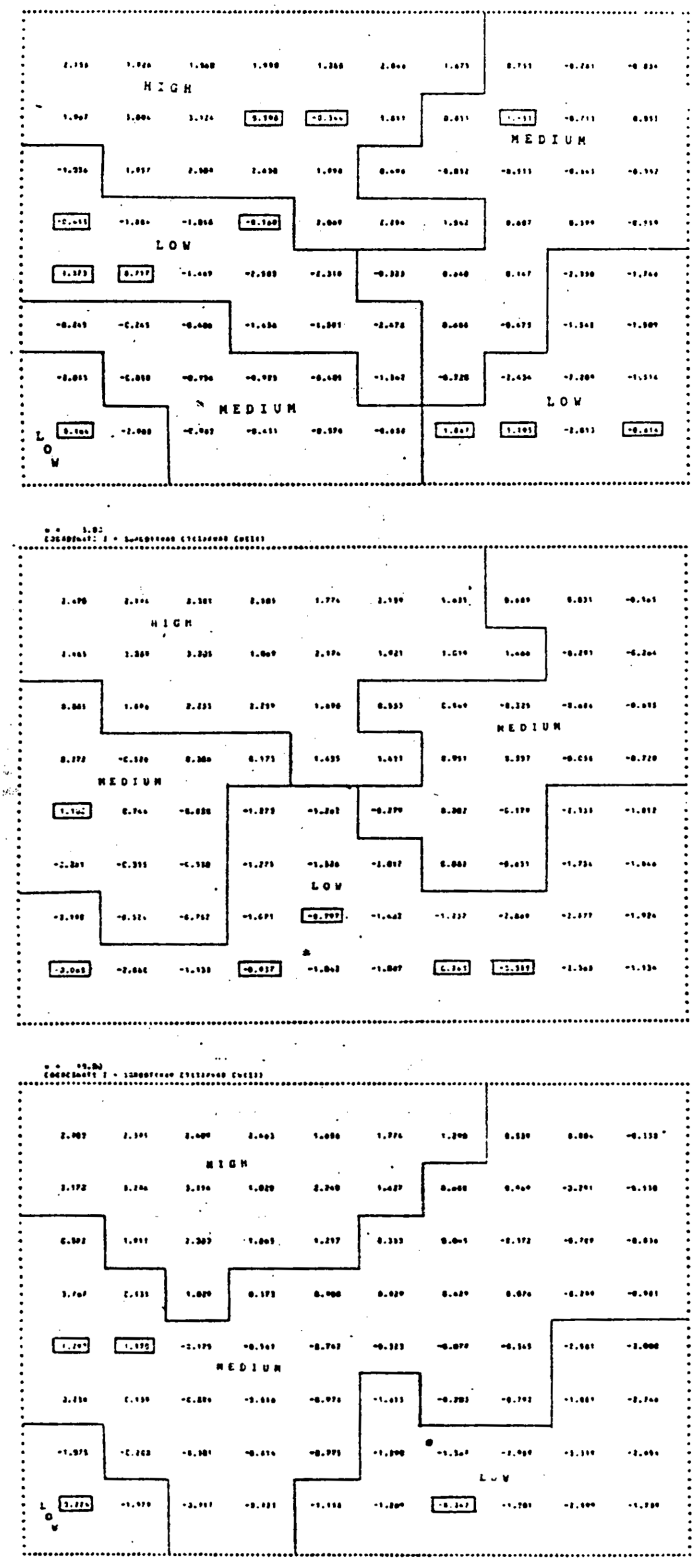
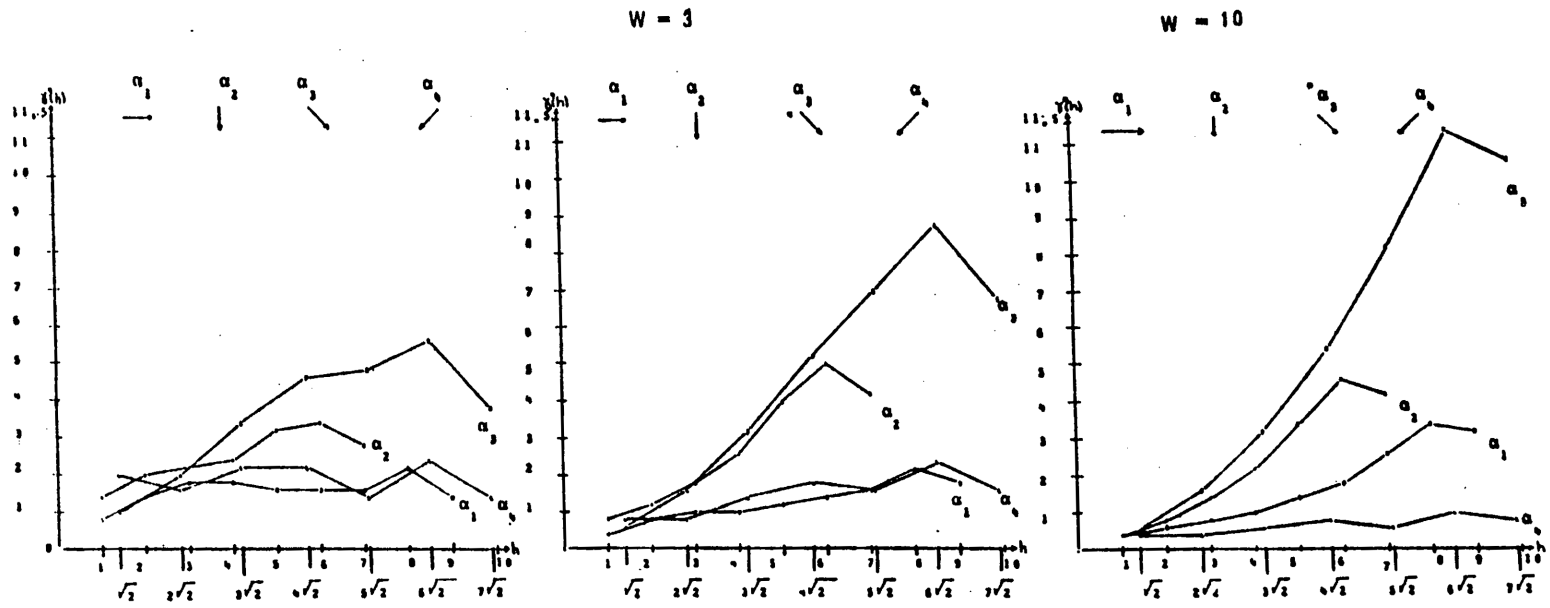


Fig. 6.1 Distribución geográfica de los valores del 1er. componente principal y los valores escalados de la 1a. coordenada principal para los 80 puntos muestrales en el área de estudio para w=3. k=10. División del área en valores altos, medianos y bajos. Los puntos aislados están encerrados en rectángulos.

Fig. 6.2 Semivariogramas experimentales para los datos no-alisados y para los datos alisados con w=3 y w=10, en las 4 direcciones principales α_1, α_2, α_3 α_4



clases formadas para w=3 y w=10 no se pueden explicar totalmente por la primera coordenada principal, esto es debido a que también hay información presente que aporta la segunda coordenada principal. Se puede apreciar una clase que se encuentra mas o menos en medio del área que aún para w=10 sigue fragmentada. Esto da evidencia de que esas dos partes de suelo separadas son de la misma clase divididas por la clase que corresponde a los valores medianos en la parte superior derecha.

Se puede decir entonces que al incrementar el valor de w utilizando el alisamiento exponencial, el número de parcelas se reduce, así como los puntos aislados y como consecuencia aumentan su tamaño promedio.

El estudio del efecto del alisamiento en la primera coordenada principal, ha aportado información objetiva acerca del valor apropiado de w. En el caso de w=3 nos da un mapa bastante razonable, aunque una de las clases esta dividida geográficamente. Si se continúa alisando (w=10) obtenemos clases de suelos contínuas, pero hay una clara evidencia de sobrealisamiento. La comparación de la varianza pepita con la varianza de las diferencias no puede decidir el mejor valor de w, pero puede al menos indicar, aún con un tamaño de muestra pequeño, un rango de valores adecuados para w.

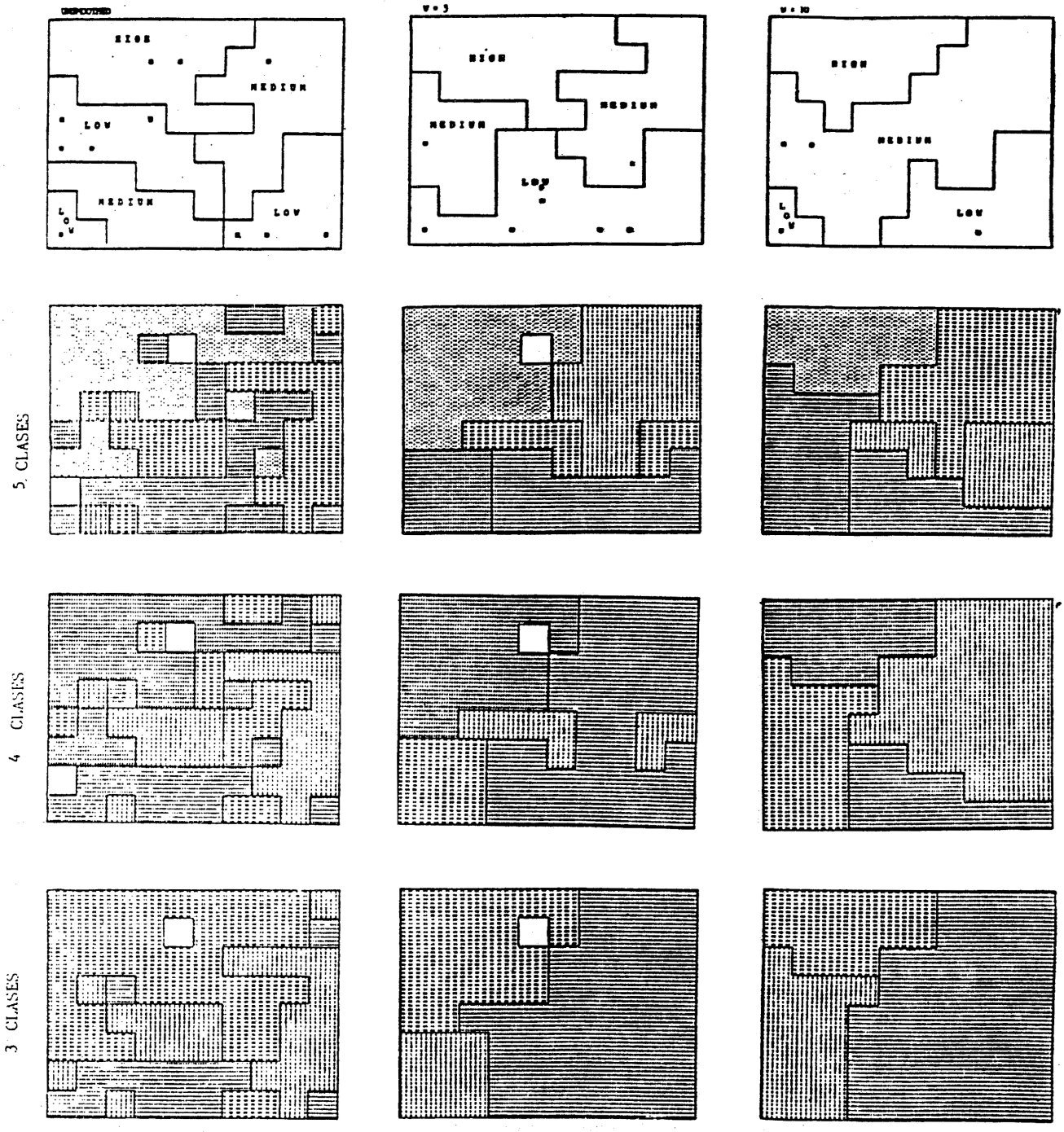
7 CONCLUSIONES

El científico en suelos se enfrenta muy seguido con decisiones difíciles de clasificación. Esta por lo general se lleva a cabo por medio de análisis de conglomerados dando mapas muy fragmentados e irreales. En este trabajo se propone un método de alisamiento basado en matrices de disimilaridad para evitar fragmentación y fronteras intrincadas. También provee un criterio para parar el alisamiento y producir un mapa que no descarta la información contenida en los datos.

Este método se puede aplicar a cualquier matriz de disimilaridad d_{ij} para producir un mapa alisado y puede ser usado con un conjunto de datos pequeño.

Una vez que se haga el alisamiento, cualquier método de análisis de conglomerados sobre la matriz de disimilaridad d^*_{ij} puede ser usado para producir una clasificación. Sin embargo, el criterio no indica "el mejor" valor para "w", pero sí descarta versiones sobrealisadas de mapas, utilizando un estimador de la varianza pepita.

Fig. 6.3 Mapas de clasificación del área de estudio, mostrando los puntos muestrales clasificados en 5, 4 y 3 clases, para los datos no-alisados y para los datos alisados con w=3 y w=10; utilizando análisis de conglomerados (ligadura completa)



8 REFERENCIAS

BURGESS, T.M. and WEBSTER, R. (1980 a). Optimal interpolation and isarithmic mapping of soil properties, I. The semivariogram and punctual kriging. J. Soil Sci.,31, 315-32.

BURGESS, T.M. and WEBSTER, R. (1980 b). Optimal interpolation and isarithmic mapping of soil properties, II. Block kriging. J. Soil Sci., 31, 333-42.

CAMPBELL, J.B. (1978). Spatial variation of the sand content and pH within single continuous delineations of two mapping units. Soil Sci. Soc. Am. J., 42, 460-64.

CAMPBELL, N.A., MULCAHY, M.J. and McARTHUR, W.M. (1970). Numerical classification of soil profiles on the basis of field morphological properties. Aust. J. Soil Res., 8, 43-58.

GOWER, J.C. (1971). A general coefficient of similarity and of its properties. Biometrics., 27, 857-71.

HAJRASULIHA, S., BANIABBASSI, N., METTHEY, J. and NIELSEN, D.R. (1980). Spatial variability of soil sampling for salinity studies in south-west Iran, Irrig. Sci., 1, 197-208.

HOLLAND, D.A.,(1969). Component analysis - An approach to the interpretation of soil data? J. Sci. Fd. Agric., 20, 26-31.

JOURNEL, A.G. and HIJBREGTS, Ch.J. (1978). Mining Geostatistics. Academic Press, London.

LANCE, G.N. and WILLIAMS, W.T. (1967a). Mixed data classificatory programs. I.Agglomerative systems. Aust. Comput. J., 1, 15-20.

MATHERON, G. (1965). Les variables regionalisees at leur estimation. Masson, Paris.

MATHERON, G. (1971). The theory of regionalized variables and its applications. Les Cahiers du Centre de Morphologie Mathematique de Fontainebleau, No. 5, ENSMP, Paris.

RAYNER, J.H. (1966). Classification of soils by numerical methods. J. Soil Sci. 17, 79-92.

RAYNER, J.H. (1969). The numerical approach to soil systematics. In: The soil Ecosystem. Systematics association publication no. 8. Editor J.G. Sheals, pp 31-39.

SNEATH, P.H.A. and SOKAL, R.R. (1973). Numerical Taxonomy. Freeman, San Francisco.

WEBSTER, R. (1977). Quantitative and Numerical Methods in Soil Classification and Survey. Clarendon Press, Oxford.

WEBSTER, R. and BURROUGH, P.A. (1972). Computer-Based Soil Mapping of small areas from sample data. Part I - Multivariate Classification and Ordination. 210-221. Part II - Classification Smoothing. 222-234. Journal of Soil Science., Vol. 23, No. 2.

WEBSTER, R. and BUTLER. B.E. (1976). Soil survey and classification studies at Ginninderra. Aust. J. Soil Res. 14, 1-24.

WEBSTER, R. and McBRATNEY, A.B. (1981). Soil segment overlap in character space and its implications for soil classification. Journal of Soil Science. 31, 133-147.

UNA APLICACION DEL ANALISIS DISCRIMINANTE SOBRE LA MADUREZ SEXUAL DE LA TRUCHA ARCO-IRIS.

Dr Gustavo J. Valencia
Laboratorio de Estadistica
Departamento de Matemáticas
Facultad de Ciencias, UNAM.

INTRODUCCION.

El cultivo intensivo de la trucha ARCO-IRIS (Salmo gairdneri) ha adquirido en México a lo largo de los últimos años, un extraordinario desarrollo debido a la aceptación de esta especie en el mercado.

El éxito de la truticultura (cultivo de la trucha) depende, en gran medida, de un manejo bio-tecnológico adecuado. La selección de reproductores resulta ser un aspecto fundamental en este proceso.

Para obtener éxito en la producción, es necesario que los reproductores se encuentren en condiciones metabólicas óptimas y que sean desovados en el momento apropiado de madurez sexual.

El análisis que se comenta en este trabajo, es parte del análisis presentado en Valencia (1986). Los datos que se analizan, son el resultado de un experimento realizado en la piscifactoria El Zarco (Km. 32.5 de la carretera México Toluca) por Rocio G Rodriguez Nieto, bajo la dirección de la M. en C. Fernanda Ruiz D. del Laboratorio de Vertebrados Acuaticos, Departamento de Biologia, Facultad de Ciencias, UNAM.

La selección de reproductores en el Zarco, se realiza de manera empirica Esto es, los piscicultores con base en los conocimientos adquiridos a través de años de práctica, realizan la selección de los individuos maduros sexualmente Una vez realizada la selección, los

individuos escogidos son tratados para que expulsen los productos sexuales (ovulos y espermatozoides). Si el individuo elegido no es sexualmente maduro, entonces no solo se le somete a una manipulación inecesaria, sino que se incrementan los costos de operación.

Un procedimiento de selección poco eficiente, implica desde un punto de vista técnico, un manejo excesivo de adultos, una posible selección a destiempo de los peces, la correspondiente pérdida de efectividad en el manejo y una elevación de los costos de operación.

Siendo la selección de reproductores, un factor muy importante para el éxito del proceso de cultivo, se pensó que en la medida de lo posible, la selección debía basarse en características cuantitativas y no cualitativas o subjetivas. Además, el estudio de los factores que influyen directamente en la selección, permitirá una mayor comprensión del problema y tal vez permita proponer algunas mejoras al procedimiento de selección.

Se consideró, que la selección debe hacerse con base en características cuantitativas morfológicas de los peces y por esto se midieron los pesos de hígado, gónada y el peso total del pez; la longitud, la altura y los índices ponderal o factor de condición, gonadosomático y hepatosomático

El objetivo que se analiza, es el de obtener un primer criterio de clasificación de madurez sexual con base en las características morfológicas antes mencionadas. Para esto, se clasificaron los peces observados, de acuerdo a la escala empírica de determinación visual, de sexo y estado de madurez gonádica; y posteriormente, se llevo a cabo un análisis discriminante

El análisis discriminante permitió explorar *las posibilidades* de obtener un criterio cuantitativo de clasificación (con una medida asociada

de efectividad) y una indicación de la importancia relativa de las variables involucradas en el criterio de selección.

LA INFORMACION.

Debido a restricciones de recursos y dado que para la realización del estudio -medición de peso de hígado y gónada, así como realizar la clasificación propuesta por Buckman (1) es indispensable sacrificar al pez, la información se baso principalmente en peces muertos.

Los ejemplares muertos, fueron recolectados -diariamente, con redes de cuchara- durante los meses de enero de 1984 a enero de 1985. Solamente en el periodo comprendido de noviembre de 1984 a enero de 1985, se obtuvó con las mismas redes, una muestra aleatoria de peces vivos.

Todos los días, durante este último periodo, se realizaba la selección de los estanques en forma aleatoria y de los estanques escogidos, se seleccionaba (tambien aletoriamente) a los peces vivos que entraron al estudio.

En el periodo de enero de 1984 a octubre del mismo año, se colectaron 395 peces muertos; y en periodo en que se colectaron peces vivos, se colectaron 150 muertos y 90 vivos

La escala de madurez propuesta por Buckman (2) es:

I.-VIRGEN -Organismos que no han alcanzado la primer madurez sexual.

Organos sexuales muy pequeños y debajo de la columna vertebral.

Testículos y ovarios transparentes e incoloros a simple vista.

1.-Se empleó la clasificación propuesta por Buckman, ya que es aceptable su uso en estudios en laboratorio para determinar la madurez sexual de peces. El método de Buckman se aplica en general para todo tipo de peces. Esto puede resultar en imprecisiones al aplicarlo específicamente sobre algunas especies de peces

2.-Puede observarse que para utilizar esta clasificación no solo debe sacrificarse el pez, sino tener experiencia en la apreciación de las características involucradas.

II.-**VIRGEN PROXIMA A MADURAR**.-Testículos y ovarios translucidos, de color rojo grisaseo, con longitudes de 3 a 8 mm. Huevos no visibles a simple vista, aunque visibles con la ayuda de una lupa.

III.-**EN PROCESO DE MADURAR**.-Testículos y ovarios opacos, rojisos con capilares sanguineos. Ocupan aproximadamente la mitad de la cavidad ventral. Los huevos son blancos, visibles a simple vista y tienen apariencia granular.

IV-**MADUROS**.-Testículos color blanco rojiso. No aparecen gotas de "lechecilla" al presionar. Ovarios anaranjados rojisos. Los huevos se observan claramente y los ovarios ocupan dos terceras partes de la cavidad ventral.

V.-**PRE-DESOVE**-Los organos sexuales llenan la cavidad ventral. Los testículos son blancos y al presionar caen gotas de "lechecilla". Los ovarios son amarillentos. Los huevos son redondos y algunos son transparentes.

VI.-**DESOVE**.- Los huevos y la lechecilla se desprenden al aplicar una presión muy ligera. La mayoria de los huevos son translucidos, con algunos opacos en el ovario. Los testículos aparecen blancos completamente.

VII.-**POST-DESOVE**-Las gónadas se aprecian flaccidas y de color sanguiñolento. Los ovarios se aprecian vacios o con pocos huevos. Los testículos pueden contener restos de esperma.

VIII.-**DESCANSO**.-Testículos y ovarios vacios y rojos. Unos pocos huevos en estado de re-absorción.

ANALISIS ESTADISTICO.

El análisis estadistico se realizó mediante el paquete SPSS, en la máquina Burroughs 6700 de la UNAM.

Las variables consideradas en el análisis serán referidas, por facilidad, en la discusión siguiente mediante los siguientes identificadores:

X_1 - madurez sexual,

X_2 - peso total,

X_3 = peso de la gónada,

X_4 - peso del higado,

X_5 = longitud,

X_6 = altura,

X_7 - indice gonadosomático,

X_8 - indice hepatosomático y

X_9 = indice ponderal.

El análisis estadistico se realizo en dos etapas: la primera, consistió en comparar las medias de las variables correspondientes a peces vivos, con las correspondientes medias de las variables en los peces muertos.

Para realizar este contraste de hipótesis, se utilizo el procedimiento presentado en Morrison (1967, pp.125-126). Este procedimiento supone normalidad e igualdad de matrices de varianza-covarianza.

La segunda etapa, consistió en un análisis discriminante.

Aunque en el trabajo original, se analizan por separado machos y hembras, en este trabajo sólo se discuten algunos de los resultados para hembras.

Las variables con las que se realizó el contraste de hipótesis fueron las variables X_2 a X_6. No se consideraron las variables X_7, X_8 y X_9 por ser estas transformaciones de las variables anteriores.

El valor observado de F es de 0.014, correspondiente a un nivel de significancia descriptivo de 0.98.

Por lo anterior, se tiene fuerte evidencia en favor de no rechazar la hipótesis de igualdad de medias entre las poblaciones de peces vivos y muertos. Estos resultados sugieren que puede continuarse el análisis sin distinguir entre peces capturados vivos y capturados muertos.

La clasificación original de los peces, se llevo a cabo utilizando la clasificación propuesta por Buckman (descrita anteriormente) y que utiliza criterios cualitativos. El análisis discriminante permitirá analizar la clasificación con base en las variables cuantitativas medidas.

Utilizando las funciones clasificantes, se procedió a revisar la clasificación resultado del análisis discriminante.

Para realizar el análisis discriminante, se supuso la igualdad de las matrices de varianza covarianza de cada uno de los grupos involucrados, así como normalidad de las variables discriminantes.

Debido al interés manifiesto por los interesados, se realizó una *selección de variables* STEPWISE, mediante el criterio de la lambda de Wilks.

Para una discusión amplia sobre el análisis discriminante, se invita al lector a consultar los trabajos de Cooley y Lohnes (1971) y de Dillon y Goldstein (1984).

En la tabla 1, se presenta un resumen del procedimiento de selección de variables para la formación de las funciones discriminantes, basado en el criterio de Wilks.

Con base en la información presentada en la tabla 2, sobre los coeficientes estandarizados de las funciones discriminantes, es posible observar que la primer función discriminante esta fuertemente determinada por los coeficientes de las variables X_3^* y X_7; por lo tanto es posible identificar a esta función discriminante, con la información de gónadas.

La función discriminante dos, esta fuertemente influenciada por la altura. La función tres, se puede asociar de esta manera con la longitud. La cuarta función discriminante esta formada principalmente por las variables longitud, altura y el índice ponderal.

En la tabla 4, se presentan los resultados correspondientes a la reclasificación de los peces, con base en las funciones clasificadoras presentadas en la tabla 3.

De la tabla 4, puede concluirse que el porcentaje de peces clasificados mediante el procedimiento obtenido a partir del análisis discriminante, sólo clasifico correctamente al 33.63% de los peces. Esto lleva a concluir, que la potencia clasificadora proporcionada por el procedimiento discriminante es pobre. Posiblemente, una razón para explicar esta pobreza, es el que algunos grupos se observaron muy poco, ya que:

0.00% del grupo 1,
1.02% del grupo 2,
13.27% del grupo 3,
5.44% del grupo 4,
14.63% del grupo 5,
24.49% del grupo 6,
32.65% del grupo 7 y
8.50% del grupo 8.

Analizando renglón a renglón la tabla 4, es posible realizar algunas observaciones interesantes.

Los grupos con menor número de observaciones, son los que presentan mayor porcentaje de observaciones mal clasificadas.

En general, todos los renglones presentan una situación mala en cuanto a clasificar. Surge entonces la pregunta: ¿las variables contempladas son importantes en el proceso clasificador?, aparentemente la información que proporcionan no es relevante o al menos se necesita de otras variables con un mayor poder discriminante.

La gráfica 1, muestra el mapa territorial. Este mapa, aunque basado sólo en las dos primeras funciones discriminantes, ilustran claramente las dificultades que enfrentaría el investigador al intentar clasificar una observación, ya que no sólo se tendrían problemas en las zonas cercanas a las fronteras, sino que hay grupos (3, 8 y 4) con áreas espaciales pequeñas y por lo tanto difíciles de identificar.

CONCLUSIONES.

El análisis discriminante, resultó en esta aplicación muy útil para explorar varias posibles modificaciones a la manera de realizar la experimentación.

Una de las principales indicaciones, es que hay que observar no sólo a los peces "grandes" y "viejos", sino tambien a los "pequeños" y "jovenes" para conocer mejor las carecterísticas de los que no son aun sexualmente maduros; y poder entonces distinguir entre los inmaduros y aquellos que ya desovaron.

Además, se consigue la indicación de que no basta con las variables observadas; y aunque la información sobre peso de gónadas e hígado es

importante y permite una mejor clasificación en los grupos 5 a 8, es conveniente analizar la inclusión de otras posibles variables discriminantes, sobre todo pensando en el tipo de información que se considera en la clasificación cualitativa y subjetiva de Buckman.

Por último, no es posible evitar el sacrificio del pez con este procedimiento y la unica posibilidad es continuar con los procedimientos empiricos de clasificación.

BIBLIOGRAFIA.

-Cooley, W W y Lohnes, P R. (1971): "Multivariate Data Analysis". Wiley, New York.

-Dillon, W.R. y Goldstein, M. (1984): "Multivariate Analysis: Methods and Applications". Wiley, New York.

-Morrison, D.F (1967): "Multivariate Statistical Methods". McGraw-Hill.

-Valencia, G. (1986): "Análisis Estadistico Sobre la Madurez Sexual en la Trucha ARCO-IRIS (Salmo gairdnieri)". Reporte No. 1, Laboratorio de Estadistica, Departamento de Matemáticas, Facultad de Ciencias, UNAM.

TABLA-1 RESUMEN DEL PROCEDIMIENTO DE SELECCION DE VARIABLES.

HEMBRAS.

PASO	ENTRA SALE	NUM DE VARIABLES EN EL MODELO	LAMBDA DE WILKS	VALOR DE F	SIGNIFICANCIA.	ETIQUETA
1	X_7	1	0.686	21.26	0.0000	INDICE GONADO-SOMATICO.
2	X_6	2	0.616	12.58	0.0000	ALTURA.
3	X_5	3	0.581	9.12	0.0000	LONGITUD.
4	X_8	4	0.558	7.32	0.0000	INDICE HEPATO-SOMATICO.
5	X_3	5	0.542	6.08	0.0000	PESO DE GONADA.
6	X_9	6	0.528	5.21	0.0000	INDICE PONDERAL.

TABLA-2 Coeficientes estandarizados de las funciones discriminantes

[redacted]

	Funciones					
VARIABLES	1	2	3	4	5	6
X_3	0.477	-0.560	-0.212	0.432	0.258	1.096
X_5	0.217	-0.602	1.021	-1.499	-0.256	-0.517
X_6	0.019	1.161	-0.263	0.968	0.068	0.241
X_7	0.605	0.410	0.255	-0.300	-0.340	-1.736
X_8	0.334	-0.040	0.218	0.339	-0.008	0.114
X_9	0.097	-0.085	-0.367	-1.027	-0.545	-0.072

TABLA-3 Coeficientes de las funciones CLASIFICADORAS

HEMBRAS

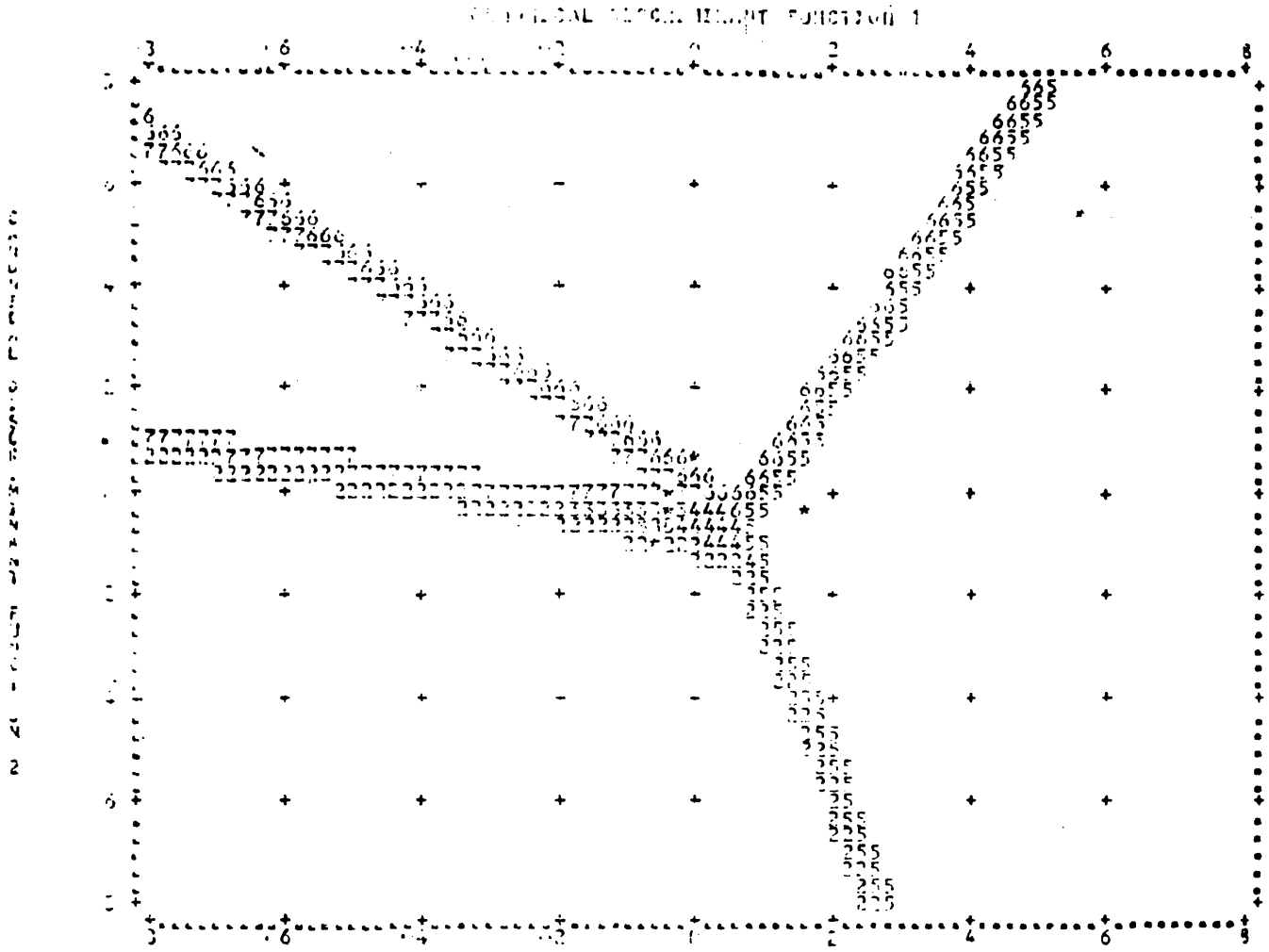
	Funciones						
VARIABLES	2	3	4	5	6	7	8
X_3	-0.046	-0.173	-0.161	-0.141	-0.179	-0.180	-0.166
X_5	2.689	2.745	2.625	2.671	2.617	2.753	2.729
X_6	-3.714	-3.635	-3.351	-3.469	-2.895	-3.376	-3.678
X_7	6.980	6.766	6.630	7.557	7.123	6.895	6.633
X_8	7.709	7.082	7.434	5.927	6.728	7.157	6.732
X_9	53.131	55.865	51.978	53.471	53.885	54.040	54.498
Const	-84.193	-89.832	-82.894	-86.317	-89.088	-90.746	-86.302

TABLA 4 Resultados de la clasificación.

REGRESAR

GRUPO ACTUAL	Num. Casos	2	3	4 *	5	6	7	8
2	3	2 66.70%	0 0.00%	1 33.30%	0 0.00%	0 0.00%	0 0.00%	0 0.00%
3	39	6 15.40%	9 23.10%	2 5.10%	0 0.00%	5 12.80%	8 20.50%	9 23.10%
4	16	6 37.50%	1 6.25%	2 12.50%	0 0.00%	1 6.25%	4 25.00%	2 12.50%
5	43	1 2.30%	3 7.00%	4 9.30%	19 44.20%	7 16.30%	5 11.60%	4 9.30%
6	72	4 5.60%	11 15.30%	10 13.90%	7 9.70%	26 36.10%	10 13.90%	4 5.60%
7	96	13 13.50%	7 7.30%	13 13.50%	0 0.00%	24 25.00%	26 27.10%	13 13.50%
8	25	5 20.00%	2 8.00%	3 12.00%	0 0.00%	1 4.00%	5 20.00%	9 36.00%

GRAFICA 1. MAPA TERRITORIAL. HEMBRAS.



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO