INSTITUTO INTERAMERICANO DE CIENCIAS AGRICOLAS – OEA
OFICINA DEL IICA EN URUGUAY
DIRECCION DE INVESTIGACIONES ECONOMICAS AGROPECUARIAS – MAP

REUNION TECNICA SOBRE TIPIFICACION DE EMPRESAS AGROPECUARIAS

Montevideo - Mayo 1977

INSTITUTO INTERAMERICANO DE CIENCIAS AGRICOLAS – OEA
OFICINA DEL IICA EN URUGUAY
DIRECCION DE INVESTIGACIONES ECONOMICAS AGROPECUARIAS – MAP

REUNION TECNICA SOBRE TIPIFICACION DE EMPRESAS AGROPECUARIAS

Editado por: Hugo E. Cohan

Montevideo – Mayo 1977

PRESENTACION

En Diciembre de 1975 publicamos tres volúmenes con los documentos presentados, comentarios y conclusiones del Seminario sobre Métodos y Problemas de Tipificación de Empresas Agropecuarias, realizado en Montevideo bajo los auspicios del IICA y la Dirección de Investigaciones Económicas Agropecuarias (DIEA), del Ministerio de Agricultura y Pesca del Uruguay.

La cantidad y calidad de trabajos discutidos, el grado de participación de profesionales directamente relacionados al tema y el notable éxito de difusión de dicha publicación -ya agotada- demostraron el acierto de haber llamado la atención hacia un problema importante para cuya solución hay técnicas disponibles, si bien son relativamente poco conocidas.

La publicación que ahora presentamos corresponde a un nuevo conjunto de trabajos realizados en Uruguay con posterioridad al Seminario mencionado. Ellos han sido elaborados en una acción de cooperación técnica de IICA con la DIEA en un intento de consolidar avances metodológicos y aplicarlos a situaciones reales. En estos esfuerzos ha resultado invalorable el apoyo metodológico ofrecido por técnicos del Centro Interamericano de Enseñanza de Estadística, cuya valiosa participación nos hacemos un deber reconocer y agradecer.

Emilio Montero
DIRECTOR
Oficina del IICA en Uruguay

SUMARIO

CAPITULO 1

Introducción del editor.

Introducción del editor

1

Hugo E. Cohan – IICA

1.1 El problema general

A lo largo de años de cooperación técnica entre el IICA y los países de la Zona Sur (Argentina, Brasil, Chile, Paraguay y Uruguay), fue surgiendo, entre varios problemas de inquietud común en el área de economía agraria, uno muy interesante. Este problema se presentó en particular en tareas relacionadas con Administración Rural y con la formulación de proyectos de desarrollo. En su máxima síntesis y llevándolo a su expresión operativa, el problema es: "¿cómo tipificar empresas agropecuarias? ".

Con la hipótesis de que distintas unidades requieren diferentes medidas de política o, lo que es similar, reaccionarán de forma diferente ante un dado conjunto de medidas generales, surge la necesidad de tipificar.

Aún dentro de economía agraria, tipificar es una acción que no requiere tomar, ni siempre toma, a la empresa como sujeto. Para distintos propósitos puede y suele considerarse que son relevantes objetos tales como unidades de producción, sistemas, regiones, zonas de suelos. No obstante ello, hay un gran número de aplicaciones posibles que requieren identificar a la empresa, por ser ella la unidad de decisión. Por ello, y pese a ser comunes a distintos casos muchos problemas de clasificación-tipificación, la discusión se identifica acá como referente a la empresa agraria.

1.2 Elementos del problema

El proceso tipificatorio es parte del análisis al que se integra y, para mejor enfocarlo, puede así descomponerse en varios puntos que requieren atención y que hacen a los objetivos, la tecnología y los recursos de este proceso. Estos puntos son:

1. objetivos del trabajo analítico, como guía de todo el proceso.
2. mecanismos estadísticos, técnicas, que faciliten la indagación sobre los datos.

3. cuerpo teórico que sugiera hipótesis sobre variables relevantes para diferenciar empresas, en función de los objetivos del trabajo.
4. una base de datos con la cual dialogar fructíferamente en las etapas de separación de grupos diferenciados, definición de casos representativos de cada grupo y validación objetiva de todo el proceso.

Sobre el primer punto, referente a objetivos del trabajo, debiera ser obvio que, desde la decisión referente a si es o no es necesario tipificar, resulta indispensable comenzar teniendo claridad en los propósitos del estudio en el que la tipificación será un instrumento. Desde un punto de vista abstracto, pudiera plantearse que tipificar es simplemente identificar poblaciones distintas para mejor estudiar el universo de empresas. Nadie negaría la importancia de tener objetivos claros, pero con un enfoque muy abstracto no habría lugar para subrayar tan definitivamente esta importancia, porque hay suficiente desarrollo de técnicas estadísticas para definir la existencia de poblaciones diferentes. La experiencia reunida en tipificación de empresas, sin embargo, indica que, por lo menos, hay dos problemas que requieren atención a los objetivos (y a la teoría de base). Estos dos problemas son:

a) las poblaciones relevantes pueden cambiar según los propósitos de diferenciarlas, y
b) dados los propósitos, la teoría y la información estadística, el analista puede ponderar ventajas y desventajas de unir o mantener separados grupos formalmente diferentes.

Sobre la tecnología a emplearse en este proceso decisorio, cabe ubicar a las técnicas estadísticas y cuasi-estadísticas con las cuales combinar recursos teóricos y de datos en un eficaz servicio a los objetivos. Estas técnicas tienen una vida propia muy relativa, en cuanto son sólo instrumentos. No obstante ello, merecen atención especial por cuanto las más aptas no son de uso común en economía agraria y el arsenal disponible era hasta hace poco desconocido para la mayoría de los usuarios potenciales. En esta área se han logrado avances importantes con la cooperación entre el IICA y los países de la zona, con apoyo del Centro Interamericano de Enseñanza de Estadística (CIENES - OEA). Incluso, la precisión de objetivos y teorías que exigen, y la base de datos necesaria para su aplicación, han hecho que el avance en la difusión de estas técnicas cumpliera una función estratégica: forzar la advertencia sobre la debilidad de recursos con la que se está operando.

Los recursos a destacarse en esta presentación son: teoría y datos. Como nota importante, debe quedar una referencia a la necesaria disponibilidad de recursos humanos capacitados para estudiar la situación de empresas agrarias. Esto debería ser parte integral de cualquier proyecto ambicioso que se inicie sobre el tema.

En cuanto a teoría, desde el muy común uso de la superficie como único criterio discriminante hasta la bastante difundida noción de relación mano de obra familiar/mano de obra contratada (tal vez combinada con la relación producción consumida/producción vendida) hay poca base conceptual y, menos aún, verificación de la misma. Predomina sí, en distintas aplicaciones, la noción de que deben reconocerse tipos diferenciados. Y existen discusiones sobre el por qué de los tipos en un dado trabajo. Pero falta un esfuerzo que integre lo ya hecho, impulse las concepciones teóricas y las someta a pruebas que permitan elaborar paradigmas científicos. A falta de esto, quien hoy tipifica debe combinar su sentido común, hipótesis generales y una bibliografía dispersa. La situación no es satisfactoria y acá se presenta claramente un frente en el que debería haber un avance en conjunto de distintos centros de investigación.

En cuanto a la situación de datos, poco cabe agregar sobre cualquier expresión de la frustración que experimenta con esto todo investigador de la realidad económico-social de América Latina. Los datos sobre empresas tienen, sin embargo, un aspecto peculiar que debe mencionarse. Este aspecto es que, a diferencia de lo que sucede con otros sujetos de indagación, es frecuente la realización de encuestas que permitirían, de ordenarse y archivarse con mecanismos de fácil acceso, una base informativa útil y en crecimiento continuo. Este problema es algo que las mecánicas para bancos de datos debieran poder resolver con relativa facilidad. La solución depende de una decisión de los países para definir para qué quieren esta información y si es razonable generarla de la manera inorgánica que hoy prevalece.

1.3 El seminario de 1975

En atención a que los métodos y problemas de la tipificación de empresas constituían un problema frecuente y común a los países, el IICA y la Dirección de Investigaciones Económicas Agropecuarias del Ministerio de Agricultura y Pesca del Uruguay, convocaron en noviembre de 1975 a un Seminario del que participaron técnicos de los distintos países de la Zona.

Los objetivos de dicho Seminario fueron:

1) Analizar en profundidad los problemas que se presentan en la tipificación de empresas agropecuarias en investigaciones en general y en elaboración de proyectos de desarrollo sectorial en particular.
2) Revisar y discutir técnicas disponibles para atender los problemas de tipificación.
3) Extraer conclusiones útiles para quienes en el futuro se enfrenten a este tema.

Observado con la perspectiva que otorgan el tiempo y experiencia adicionales, sobresale el logro de los objetivos segundo y tercero. En efecto, la revisión y discusión de técnicas ocupó la atención preferente de los participantes y ha resultado en una acción de gran difusión e impacto en los países miembros del Instituto. Las conclusiones generadas en el Seminario exponen con claridad recomendaciones absolutamente válidas. Si acaso, la gran atracción de técnicas novedosas ha resultado en un desbalance en relación a los esfuerzos necesarios para mejorar el cuerpo teórico e impulsar el mejoramiento de los datos. Más aún, el entusiasmo inicial por emplear estas técnicas en ocasiones superó a las advertencias y recomendaciones que los mismos participantes dejaron por escrito.

La gran acogida que tuvo la publicación emanada de este Seminario, rápidamente agotada, y las numerosas aplicaciones que sobre el tema se generaron en nuestros países, hicieron oportuno intentar una evaluación de lo logrado. Esto se intentó en la Reunión Técnica que resultó en este volumen.

1.4 La reunión técnica de 1977

El IICA y la Dirección de Investigaciones Económicas Agropecuarias organizaron esta reunión en mayo de 1977. Nuevamente se contó con la participación de personal del Centro Interamericano de Enseñanza de la Estadística.

La reunión se organizó en torno a tres áreas de exposiciones. A saber:

1) profundización de conocimiento sobre las técnicas de tipificación propuestas en el Seminario de 1975.
2) análisis de trabajos aplicados.
3) presentación teórica de propuestas de tipificación.

La Reunión puso énfasis en las dos primeras áreas temáticas y los correspondientes trabajos se presentan en este volumen. Las propuestas de tipificación presentadas por Martín Piñeiro (proyecto sobre creación y adopción de tecnología), Carlos Pérez Arrarte (proyecto de estudio ganadero) y Roberto Casás-Beatriz Licio (proyecto de regionalización agropecuaria) permitieron un primer intento de discusión teórica seria. El difícil armado de una presentación orgánica de lo tratado sobre estos proyectos con variante grado de desarrollo analítico, impidió que se editaran las correspondientes discusiones.

Para profundizar el conocimiento de las técnicas propuestas en 1975 se contó con los trabajos de Guillermo Artigue (componentes principales y factorial) y de Alfredo Alonso (clustering), así como con los aportes de Pedro Ferreira y Mario Kaminsky.

El documento de Artigue plantea como opción operativa para el análisis factorial, un modelo basado en la extracción de componentes principales. Esto es lo que alguna literatura especializada (Cattell, citado por Artigue) también denomina "modelo cerrado" de análisis factorial. No todos los expertos aceptan que esta sea una forma válida de hacer análisis factorial y, en particular, se registraron durante la reunión discusiones acerca de si es o no válido mencionar dos formas de extraer factores (la "pura" y la "basada en componentes principales"). El editor no está en condiciones de proponer un dictamen fundado sobre esta controversia. Pero cabe sí indicar que el aporte de Artigue y la consiguiente controversia permitieron a los participantes de la reunión tomar una clara idea de qué está involucrado en las operaciones más usuales, las que se han basado en la extracción de componentes, la retención de los más significativos y la denominación de la varianza no explicada como "factor común", por similitud al análisis factorial.

Alfredo Alonso presentó en esta primera parte de la reunión una amplia exposición sobre las técnicas de conglomeración para las cuales están disponibles en Uruguay programas de cómputo desarrollados para el IICA. Con menor extensión, también trató el tema de técnicas de verificación de resultados. Esto se había previsto como aspecto importante del temario y fue enfatizado por Ferreira y Kaminsky.

Ferreira centró su atención en la conveniencia de controlar los grupos que se formen y en métodos disponibles para evaluar conglomerados. Con palabras muy similares a las por él empleadas, sugirió pruebas destinadas a cuidar que los agrupamientos resultantes fueran naturales del universo en estudio y no imposiciones de un dado, infeliz, cruzamiento entre un mecanismo de conglomerar y unos datos.

Kaminsky se centró en teoría y ejemplos de cómo validar y cómo no validar tipificaciones. Su énfasis, combinado con aplicación de técnicas propuestas en el aporte de Ferreira, es en sí misma una evaluación crítica de nuestro primer año de experiencias en tipificación con arsenal estadístico y cuasi-estadístico. Pese a que el posible problema estaba previsto en las conclusiones del Seminario de 1975, la usual tensión entre necesidad operativa y requisitos científicos puede resultar en un mal uso-abuso de tecnologías. Si, como parece, puede prevalecer la tentación (tal vez inconsciente) de usar el arsenal formal disponible como justificación de cualquier causa, esta contribución-advertencia de Kaminsky debe ser particularmente bienvenida. El peligro es común a cualquier discurso científico y la difusión no meditada de técnicas cuantitativas puede contribuir a su generalización. Eso estaba claro en las conclusiones del Seminario de 1975 y, no obstante, es un punto en el que cabe insistir.

Los trabajos aplicados de Dabezies-Sarroca (ganadería), Heraclio Pérez (lechería) y Alfredo Alonso (regionalización) permitieron una útil discusión y evaluación de experiencia. El apoyo de Artigue, Ferreira y Kaminsky permitió a los autores hacer y corregir sus documentos. Las versiones que acá se presentan recogen algunas opiniones de los críticos. Queda a los lectores de este volumen el decidir si en la actual versión satisfacen o no la meta de mayor objetividad científica que guiaron esta acción iniciada en 1975.

1.5 Conclusiones del editor

La experiencia reunida en un año de aplicación de nuevas técnicas fue exitosa. Sin perjuicio de lo mucho que aún falta por desarrollarlas y difundirlas convenientemente, lo logrado en este campo es lo suficiente como para pensar en un desbalance del proceso de tipificación.

Nuevos desarrollos debieran en lo futuro provenir especialmente del área de los recursos teóricos y de datos sobre los que se opera al tipificar. Como aspecto parcial que es, la tipificación no puede avanzar mucho más por las mismas causas que limitan nuestra capacidad general de análisis sobre empresas.

En las distintas acciones que emprenden en economía agraria el IICA y los países, el objetivo es el de comprender mejor la realidad y saber cómo adaptarla a las aspiraciones nacionales. La unidad empresa agraria es central en la realidad sectorial que interesa. Y así como las técnicas de tipificación han resultado

útiles también por su rol estratégico en definir bloqueos a su mejor uso, es de esperar que avances en las débiles áreas de teoría y datos permitan jugar al proceso de tipificación de empresas un rol estratégico. Ese rol es el de mejorar la comprensión de la realidad y la capacidad para equilibrar hechos y aspiraciones.

La publicación de estos trabajos tiene el objetivo de convocar a una mayor audiencia a cooperar en el difícil trabajo aún pendiente.

Hugo E. Cohan

CAPITULO 2

Componentes principales y análisis factorial. Su naturaleza y sus posibilidades en tipificación.

Componentes principales y análisis factorial. Su naturaleza y sus posibilidades en tipificación.

2

Guillermo Artigue – DIEA

2.1 Introducción

En el Seminario sobre Métodos y Problemas en la Tipificación de Empresas Agropecuarias, realizado en Montevideo en noviembre de 1975*, se presentaron como técnicas útiles para la Tipificación de Empresas Agropecuarias los temas "Componentes Principales" y "Análisis Factorial", entre otras.

En el presente trabajo se desarrollan un poco más estos temas, con los propósitos de facilitar la comprensión de sus bases teóricas y enfatizar tanto el rol de ellos en la formación de grupos de empresas, como en el análisis de factores no observables responsables de los agrupamientos.

En el citado Seminario quedó claro que, en el supuesto de poseer información básica procesable de las empresas bajo estudio, estos temas estadísticos aportan grandes posibilidades a un proceso tipificatorio de las mismas. La posibilidad de reducir la información a un conjunto pequeño de variables o factores, el análisis de éstos y la conglomeración de las empresas en función de características subyacentes, aparecen como etapas de un proceso estadístico que debe conducir a una tipificación objetiva.

A estos efectos se presentará el tema "Componentes Principales", porque permite resumir en un conjunto pequeño de variables no correlacionadas, casi toda la información proveniente de un relevamiento parcial (muestra) o completo, de un universo de empresas. Mediante el cruce de cortes realizados en estas variables, es posible estratificar el universo o, mejor aún, realizar un Análisis de Conglomeración utilizando una matriz de similaridades entre empresas, extraída de los valores que toman las Componentes Principales para cada empresa.

El tema "Análisis Factorial" será presentado en las dos versiones siguientes:

A. Análisis Factorial con factores comunes y factores específicos.

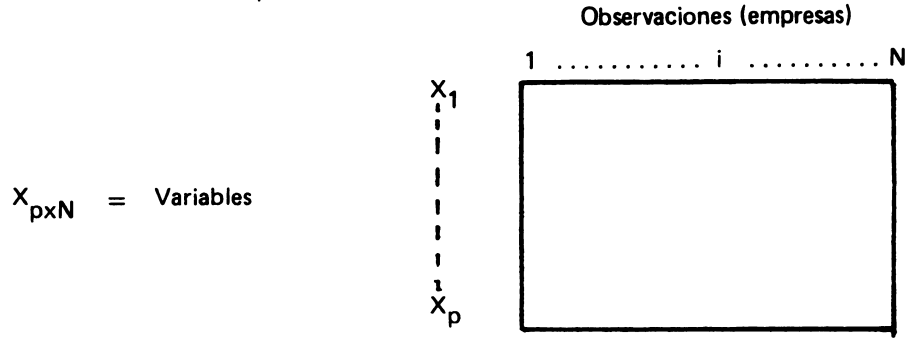
B. Análisis Factorial en Componentes Principales.

---

* Véase el trabajo de P. Ferreira, en (7), Vol. 1, Capítulo 5.

Se analizarán las posibilidades y limitaciones de ambos Análisis, tanto en los aspectos teóricos como metodológicos de sus aportes a un proceso tipificatorio de empresas.

La información básica que se supone disponible de partida para las presentaciones que se harán, puede provenir del relevamiento total de un conjunto relativamente chico de empresas de interés en sí mismas, o de una muestra representativa de un universo suficientemente grande como para recomendar su muestreo. En cualquiera de los dos casos, la información se presenta en la forma de una matriz X de observaciones, en donde los p elementos del i-ésimo vector columna (observación), corresponden a los p atributos (o variables respuestas) obtenidos de la i-ésima empresa:



Esta notación debe tenerse en cuenta en todo lo que sigue.

Dado que en lo sucesivo se trabajará con varianzas, covarianzas y correlaciones, se supone que las variables respuesta han sido "centradas". Esto significa que los elementos X_{ij} de la matriz X se definen como sigue:

$$X_{ij} = v.o - \overline{v.o}$$

donde v.o significa "valor original" y $\overline{v.o}$ es "promedio de valores originales", correspondiendo ambos a una misma variable. Nótese que las varianzas, covarianzas y correlaciones calculadas con valores originales o con valores centrados, son las mismas*.

El conjunto de observaciones representado en un espacio de p dimensiones, proporciona una nube de N puntos (cada punto corresponde a una empresa). Este conjunto de puntos tiene el origen de los ejes en su centro de gravedad.

En un espacio de N dimensiones puede representarse no ya las empresas sino cada una de las p variables, obteniéndose también una nube de p puntos.

El análisis puede así privilegiar a las N observaciones o a las p variables. Por razones de presentación en este trabajo se pone énfasis en las N observaciones (empresas).

2.2 Componentes principales

2.2.1. Introducción al Tema

Componentes Principales puede concebirse como una transformación de coordenadas de las observaciones originales. Mediante esta transformación, el sistema de ejes original se gira hasta ubicarse en las direcciones de la nube de puntos (representativa del conjunto de empresas) que presentan mayor dispersión.

---

* Véase (9), pág. 108.

En un espacio de dos dimensiones, si se denominan por X_1 y X_2 a las variables respuestas de cada empresa, y por Y_1 e Y_2 a las nuevas variables (Componentes Principales), puede interpretarse lo expuesto mediante las siguientes gráficas:

– **Las componentes como un nuevo sistema de ejes**

La componente principal Y_1 se ubica en la dirección donde claramente hay mayor dispersión de todas las direcciones posibles. La segunda componente principal se ubica en la dirección perpendicular a la primera.

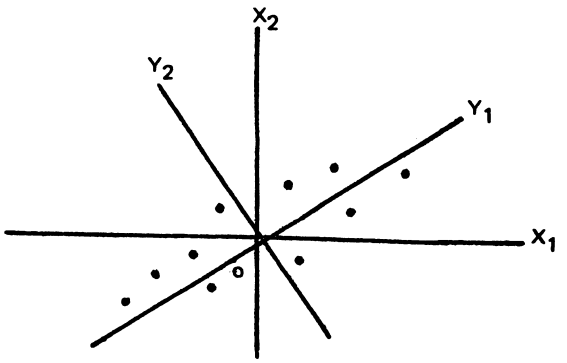


Figura 1. Las componentes proporcionan un nuevo sistema de ejes.

– **Proyección de las observaciones sobre los dos sistemas de ejes.**

Las coordenadas de cada empresa se obtienen proyectando en forma perpendicular sobre los ejes. Los valores orlginales se obtienen proyectando sobre X_1 y X_2 y los valores que adoptan las Componentes Principales, para esa empresa, se obtienen proyectando sobre las nuevas direcciones Y_1 e Y_2.

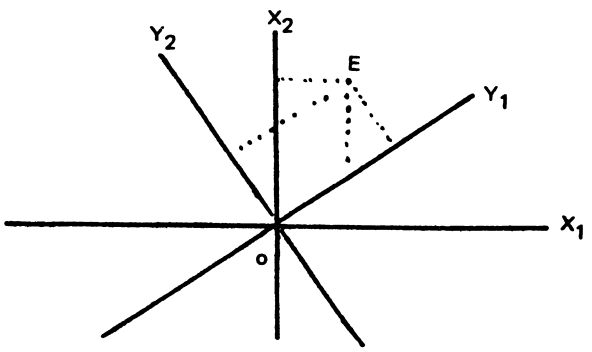


Figura 2. Valores originales y en relación a componentes, de una empresa.

Las proyecciones sobre Y_1 presentan la máxima dispersión posible de todas las que pueden obtenerse proyectando las empresas sobre ejes que pasen por el origen. Tal dispersión es medible en términos de varianza y en función de ella se elegirá la dirección conveniente.

La suma de las varianzas de las componentes principales es igual a la suma de las varianzas de las variables originales ("varianza total del sistema de N empresas"). Por lo tanto, puede calcularse la contribu-

ción de cada variable Componente Principal a la varianza total del sistema, dividiendo la varianza de cada Componente entre la varianza total.

Es posible que unas pocas Componentes Principales aporten un porcentaje importante a la varianza total. En ese caso, sustituyendo las variables originales por estas pocas nuevas variables, se pierde poca información y se gana mucho en simplicidad.

Así surge claramente un posible uso de esta técnica en un proceso tipificatorio: la reducción del conjunto de variables originales, haciendo manejable un volumen de información que de otra manera podría ser sumamente difícil de procesar.

Como utilización conexa con la que nos interesa fundamentalmente, es destacable que la técnica ha sido utilizada en encuestas de objetivos múltiples*. Para ello, se cruzaron cortes realizados sobre las primeras componentes (las cuales acumulaban un porcentaje importante de la varianza total), definiéndose de esta manera "estratos".

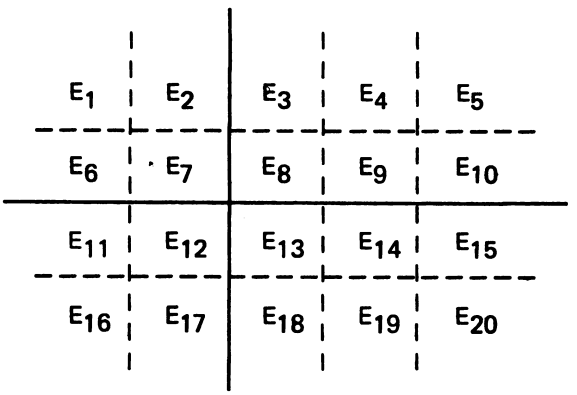


Figura 3

La Figura 3 representa una situación en la que se conservaron las primeras dos Componentes Principales. Se cruzaron cortes en ambas, definiéndose veinte estratos. Si de cada estrato se elige al azar un representante, se tiene así una muestra de tamaño veinte.

Otras técnicas de Conglomeración, analizadas en el trabajo que presenta A. Alonso a esta reunión, se basan en matrices de similaridades entre empresas. Estas similaridades suelen construirse en función de las variables originales. En caso de ser numerosas estas variables, ellas pueden reducirse primero a un conjunto pequeño de Componentes Principales y con estas construirse la matriz de similaridades.

En cualquiera de los casos analizados, el tema Componentes Principales brinda su aporte simplificando el volumen de información mediante la reducción de un conjunto numeroso de variables a un número pequeño de nuevas variables, con una pérdida de información controlable por el número de componentes principales retenidas (en función del porcentaje de la varianza total que ellas explican).

---

* Véase por ejemplo (6).

2.2.2. Conceptos teóricos básicos

Las variables Componentes Principales se calculan en forma sucesiva. Dado que su cálculo consiste en una transformación de coordenadas a través de un giro de ejes, estas nuevas variables se expresan en forma lineal respecto de las originales.

La primer Componente Principal Y_1 se define como la expresión lineal de las variables originales que presenta mayor varianza

$$Y_1 = a_{11} X_1 + \ldots\ldots + a_{1p} X_p$$

Exigiendo que el vector $a_1 = (a_{11}, \ldots . a_{1p})$ tenga longitud uno (normalizado), se logra determinar completamente* los números $a_{1j}, j = 1,p$.

Este vector a_1 determina la dirección del nuevo primer eje.

La segunda Componente Principal Y_2 es la segunda expresión lineal de las variables originales, con mayor varianza:

$$Y_2 = a_{21} X_1 + \ldots\ldots + a_{2p} X_p$$

en la que el vector $a_2 = (a_{21} \ldots . a_{2p})$ tiene longitud uno **y es perpendicular al vector a_1** (que determina a la Primer Componente). Esta condición de perpendicularidad entre los vectores a_1 y a_2, asegura que las variables Y_1 y Y_2 no estén correlacionadas**. Por lo expuesto, Y_2 se obtiene buscando de entre las direcciones perpendiculares a Y_1, aquélla en la cual la nube de puntos presenta mayor dispersión.

De esta forma, se van extrayendo en forma sucesiva las p variables Componentes Principales, exigiendo linealidad respecto de las variables originales, máxima varianza y perpendicularidad con las anteriormente extraídas.

Este proceso de "extracción" de las Componentes Principales se sistematiza utilizando algunos conceptos matemáticos que a continuación se repasan brevemente.

2.2.3. Valores y vectores propios de una matriz

Toda matriz cuadrada A puede interpretarse como una transformación de **un espacio de vectores,** en sí mismo. En esta transformación, todo vector se transforma en otro vector del mismo espacio: el vector μ de la figura adjunta se transforma en el vector v.

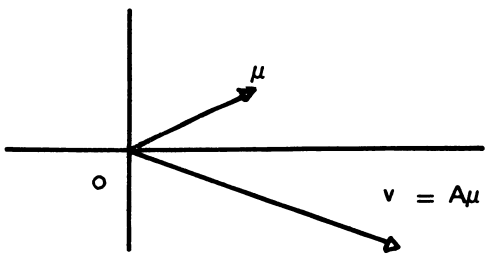


Figura 4. La matriz cuadrada A transforma el vector μ en el vector v, dentro del mismo espacio.

---

* Véase (10) pág. 224

** Idem (10) pág. 228

Como ejemplo de este concepto, considérese:

$$A = \begin{bmatrix} 2 & 0 \\ 2 & -1 \end{bmatrix}$$

Cualquier vector en el espacio bi-dimensional (por ejemplo: $\mu = (1,2)$, será transformado en otro vector v_1 preservando el espacio de dos dimensiones. Así:

$$v = A\mu$$

$$v = \begin{bmatrix} 2 & 0 \\ 2 & -1 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

Gráficamente, el ejemplo resulta en:

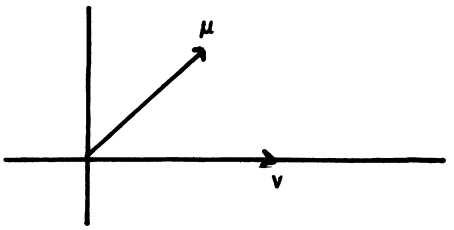
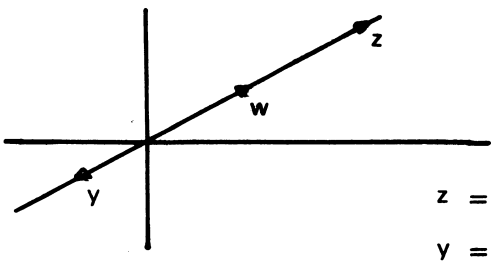


Figura 5. Ejemplo de transformación en el mismo espacio.

Sea w un vector cualquiera. Todo vector perteneciente a la misma recta que w puede obtenerse multiplicando a éste por un número real conveniente. Gráficamente:



$z = a.w \quad \text{con } a > 0$

$y = b.w \quad \text{con } b < 0$

Figura 6. El escalar "a" expande o contrae al vector w, preservando su ángulo.

En base a estas nociones podemos introducir los conceptos de valores propios y vectores propios. Estas nociones, que también aparecen en la literatura como valores y vectores "eigen" o como valores y vectores "característicos", se emplean mucho en los cálculos y desarrollos teóricos del tema que nos interesa.

Si se verifica que $A\mu = a\mu$

donde A es una matriz cuadrada n x n, μ es un vector n - dimensional y "a" un número real, entonces:

1. μ es un vector propio de A, y
2. él está asociado al valor propio a.

En el caso de la matriz del ejemplo precedente:

$$A = \begin{bmatrix} 2 & 0 \\ 2 & -1 \end{bmatrix}$$

tenemos que sus valores propios son 2 y –1.

Para obtener un vector propio, podemos resolver:

$$\begin{bmatrix} 2 & 0 \\ 2 & -1 \end{bmatrix} \mu = 2\mu$$

El resultante sistema de ecuaciones puede resolverse con el vector $a_1 = \binom{3}{2}$ o con cualquier expansión o contracción de él, del tipo ka_1 donde k es un número real.

A su vez, al valor propio –1 de la matriz A, le corresponde el vector $a_2 = \binom{0}{1}$ o cualquier transformación de él que no altere su dirección.

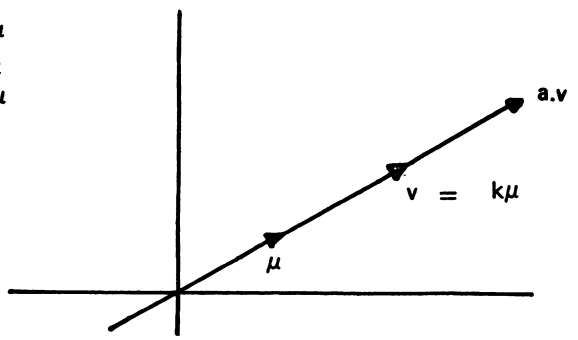
La indicación de que un vector cualquiera en el correspondiente espacio (sea μ), es un vector propio de A significa que él responde a la condición ya indicada:

$$A\mu = a\mu$$

Esta condición, a su vez, significa que, mediante la matriz A, el vector μ se transforma en otro vector **perteneciente a la misma recta (o dirección).**

Además, cualquier otro vector de esta dirección, también es propio y su transformado pertenece a la misma dirección. En efecto, si $v = k\mu$ es otro vector de la dirección del vector μ entonces:

$$\begin{aligned} Av &= Ak\mu \\ &= kA\mu \\ &= ka.\mu \\ &= a.k\mu \\ &= a.v \end{aligned}$$



Se dice que los vectores propios de una matriz determinan "direcciones propias" de la transformación asociada a la matriz, entendiéndose por "direcciones propias" aquellas direcciones que no son afectadas por la transformación (se transforman en sí mismas).

En nuestro ejemplo, las direcciones propias de la transformación que define la matriz A se identifican con las rectas a las cuales pertenecen a_1 y a_2. Cualquier vector de estas direcciones, al transformarse, se mantiene en la misma dirección.

Estos conceptos, a los que se presentó con algún detalle, sirven para calcular las variables Componentes Principales en virtud de los resultados que se exponen a continuación.

Sea S la matriz de varianzas y covarianzas de las variables respuestas de las empresas. La matriz S tiene asociados p valores propios*, todos mayores o iguales a cero, y los p vectores propios asociados a estos valores propios, determinan direcciones perpendiculares en el espacio de las respuestas. Además, estas direcciones propias perpendiculares, determinan a las variables Componentes Principales y los valores propios son sus respectivas varianzas.

Es decir: "Los vectores a_i, $i = \overline{1,p}$ que determinan las direcciones de las variables Componentes Principales Y_i, son los vectores propios de la matriz S de varianzas y covarianzas de las variables originales, y los valores propios l_i $i = \overline{1,p}$ de la matriz S son las respectivas varianzas de las variables Y_i.
Se define la varianza total del sistema como "la suma de las varianzas de las variables originales".

$$V = \sum_{i=1}^{p} \text{Var } (X_i)$$

El valor V es invariante (no cambia) al realizar la transformación de coordenadas que conduce a las Componentes Principales**. Por lo tanto:

$$V = \sum_{i=1}^{p} \text{Var } (Y_i)$$

Si y es un vector cuyos elementos son las variables Y_i, $i = \overline{1,p}$, entonces, en virtud de la perpendicularidad de los vectores de longitud uno que definen a las Componentes Principales, se demuestra*** que la matriz de varianzas y covarianzas del vector y es:

$$\text{COV } (y) = L = \begin{bmatrix} l_1 & & 0 \\ & \ddots & \\ 0 & & l_p \end{bmatrix}$$

Es decir, L es una matriz cuyos elementos de la diagonal principal son los valores propios de la matriz S, o sea las varianzas de Y_i, $i = \overline{1,p}$ y el resto de sus elementos son ceros.

---

* Véase (10), pág. 224

** Véase (5), pág. 7

*** Véase (4), pág. 206

Se define "traza" de una matriz a la suma de los elementos de la diagonal principal. Por consiguiente, la varianza total del sistema se expresa también por:

$$V = \text{traza } (S)$$

$$= \sum_{i=1}^{p} l_i$$

$$= \text{traza } (L)$$

La importancia relativa de la contribución explicativa que hace la i-ésima Componente Principal puede así medirse por:

$$\frac{l_i}{V}$$

donde l_i es el i-ésimo valor propio más grande de S.

Estos conceptos y resultados matemáticos permiten efectuar la "extracción de las Componentes Principales en forma sistemática y ordenada. Mediante un programa de computadora que calcule la matriz S de varianzas y covarianzas de las observaciones originales, que determine sus valores propios en forma ordenada, de mayor a menor, y los vectores propios asociados de longitud uno, se pueden extraer las Componentes Principales en orden de importancia de acuerdo a su contribución explicativa.

En lo expuesto se ha trabajado con la matriz S de varianzas y covarianzas, calculada a partir de los valores originales. Pero podría objetarse que estos usualmente se miden en unidades diferentes (hectáreas, cabezas de ganado, caballos de fuerza, etc.). Por lo tanto, los resultados que se obtienen son difíciles de interpretar. Para resolver este problema se puede estandarizar* las variables de partida. De hacerse esto, la matriz S se convierte en la matriz R de correlaciones originales y todo lo expuesto puede desarrollarse en función de ella. En estas condiciones, la varianza total del sistema es igual a la traza de R, o sea igual a p, número de variables originales. La varianza de la i-ésima Componente Principal es l_i, i-ésimo valor propio mayor de R. Luego, la importancia relativa de la contribución de Y_i se mide por: $\frac{l_i}{P}$

Se dispone en estos momentos en el IICA de un programa de computadora desarrollado por IBM, en el cual se determina en primer lugar la matriz de correlaciones R a partir de las observaciones originales. Esto es equivalente a ponderar las variables originales por los inversos de sus respectivos desvíos típicos y luego calcular la matriz de covarianzas de las variables ponderadas. Se determinan a continuación los valores propios l_i de la matriz R en forma ordenada, de mayor a menor. Luego, para cada l_i, se calcula el vector propio asociado de longitud uno. El vector propio correspondiente a l_1 (mayor valor propio de R), determina la dirección de la primer Componente Principal, la que aporta mayor cantidad de varianza a la varianza total del sistema. La cantidad l_1/p mide su importancia explicativa. De esta manera y en forma ordenada, se extraen todas las Componentes Principales.

La salida impresa de este programa es optativa, en el sentido de que se imprimen sólo aquellos vectores propios que corresponden a valores propios mayores que un valor prefijado por el usuario. Este puede decidir que aquellas Componentes que aportan menos de un 1 % a la varianza total no contribuyen en forma sustancial y puede fijar el valor límite para la impresión mediante la ecuación:

---

* Estandarizar una variable significa centrarla (restarle su promedio) y luego dividirla entre su desvío típico: $R_i = \frac{X_i - \bar{X}'}{S}$ El conjunto así corregido tendrá media igual a cero y varianza de 1.

$$\frac{l_i}{p} = .01$$

donde p es el número de variables.

De esta ecuación se puede despejar $l_i = .01\ p$. Este valor puede ser el que se ingresa al programa. Si se desearan imprimir todos los valores y vectores propios, debería ingresarse un cero o cualquier número negativo.

Ejemplo:

El siguiente es un ejemplo que pretende únicamente presentar la salida impresa del programa de que se dispone en el IICA, Uruguay, para Componentes Principales. Dado que se trata de la primera parte de un programa más amplio de Análisis Factorial, su encabezamiento comienza con: FACTOR ANALYSIS .... (Nombre Identificatorio).

Se compiló en tarjeta perforada la siguiente matriz de observaciones:

		EMPRESAS						
VARIABLES:		1	2	3	4	5	6	7
X_1	–	200	500	100	800	300	1.000	400
X_2	–	150	420	70	500	200	600	350
X_3	–	50	30	10	100	80	50	20
X_4	–	2	3	1	4	4	4	3
X_5	–	10	21	8	45	30	40	15

Con esta información, se corrió el programa. La salida impresa presenta, como control, el número de casos o empresas y el número de variables por empresa. Para cada variable se imprimen, por filas, sus promedios y desvíos típicos. A continuación se imprime la matriz de correlaciones generada con las variables respuestas e, inmediatamente, una fila con los "eigenvalues" o valores propios de la matriz R. Se calculan los porcentajes acumulativos de los valores propios sobre la varianza total y se imprimen en una fila. Finalmente, aparecen impresos los Eigenvectors o vectores propios de la matriz R, correspondientes a los valores propios impresos más arriba y en el mismo orden. Es decir, se imprime la matriz A^T.

La salida de computadora parcial correspondiente a este ejemplo, se presenta a continuación.

La matriz A^T, cuyas filas con los vectores propios a_i $(i = \overline{1,p})$ de longitud uno, que surgen de R, define la transformación que convierte a los ejes originales en los nuevos ejes o direcciones de las Componentes Principales. El supra-índice T de la matriz A significa "traspuesta de la matriz A". Por lo tanto la matriz A tiene por columnas las filas de A^T, o sea: los vectores propios de R.

Como los vectores a_i $(i = \overline{1,p})$ son perpendiculares entre sí, y además tienen longitud uno, resulta que la matriz A^T es una matriz ortogonal* y por ende su matriz traspuesta A es también su matriz inversa.

---

* Véase Finkbeiner (4), pág. 198

FACTOR ANALYSIS ..... ARTIGU

NO. OF CASES 7
NO. OF VARIABLES 5

MEANS

471.42847	327.14282	50.00000	3.00000	24.14285

STANDARD DEVIATIONS

325.13721	194.56909	31.09126	1.15470	14.57656

CORRELATION COEFFICIENTS

ROW 1				
1.00000	0.96801	0.44515	0.75468	0.86610
ROW 2				
0.96801	1.00000	0.35265	0.74925	0.78409
ROW 3				
0.44515	0.35265	1.00000	0.74278	0.77596
ROW 4				
0.75468	0.74925	0.74278	1.00000	0.88128
ROW 5				
0.86610	0.78409	0.77596	0.88128	1.00000

EIGENVALUES

3.95947	0.81629	0.15672	0.05714	0.01035

CUMULATIVE PERCENTAGE OF EIGENVALUES

0.79189	0.95515	0.98650	0.99793	1.00000

EIGENVECTORS

VECTOR 1				
0.46180	0.44232	0.36731	0.46718	0.48776
VECTOR 2				
−0.40388	−0.50412	0.73336	0.17474	0.11992
VECTOR 3				
−0.33034	0.06044	−0.28601	0.83571	−0.32711
VECTOR 4				
−0.00698	−0.49878	−0.49382	0.13783	0.69879
VECTOR 5				
−0.71724	0.54569	−0.03993	−0.18384	0.39037

Si se denota por x al vector de las variables originales y por y al vector de las Componentes Principales, la transformación de coordenadas que convierte x en y se expresa, en notación matricial, por:

$$y = A^T \cdot x$$

Es decir, la primera fila de A^T (primer vector propio de R) "multiplicada" por el vector de respuestas x proporciona Y_1. La segunda fila de A^T "multiplicada" por x proporciona Y_2 y así sucesivamente hasta obtener Y_p.

Si el analista retiene solamente m componentes ($m < p$) para usos tipificatorios, son útiles solamente las primeras m filas de A^T, es decir: los primeros m vectores propios de R. Multiplicando a éstos por el vector observación estandarizado correspondiente a una empresa, pueden evaluarse las respectivas Componentes Principales para esa empresa.

Ejemplo:

Para evaluar la primer Componente Principal correspondiente a la empresa número 3 del ejemplo anterior, debe tenerse en cuenta que se trabajó con la matriz de correlaciones. Por lo tanto, las respuestas deben estandarizarse. El vector observación estandarizado correspondiente a la empresa 3 es:

$$\begin{aligned}
(100 - 471.43) &\div 325.14 = -1.14 \\
(70 - 327.14) &\div 194.57 = -1.32 \\
(10 - 50) &\div 31.09 = -1.29 \\
(1 - 3) &\div 1.15 = -1.74 \\
(8 - 24.14) &\div 14.58 = -1.11
\end{aligned}$$

A continuación se efectúan los productos de los elementos correspondientes de este vector observación y del primer vector propio, sumando luego los resultados. El valor final es el valor que adopta la primer Componente Principal para esta empresa:

$$Y_1 = (-1.14) \times 0.46 + (-1.32) \times 0.44 + (-1.29) \times 0.37 + (-1.74) \times 0.47 + (-1.11) \times 0.49 = -2.94$$

2.2.4. Resumen del Capítulo

Componentes Principales es un tema estadístico que aporta soluciones a muchos tipos de problemas. Dada su característica de resumir una información muy vasta, puede utilizarse como técnica exploratoria o como técnica clasificatoria. Mediante su uso se resume la información proporcionada por un conjunto numeroso de variables a un conjunto menor de variables independientes. Existe evidentemente una pérdida de información, pero esta pérdida es controlable por la cantidad de Componentes Principales extraídas.

Si para cada empresa se evalúan las Componentes Principales, es posible clasificar las empresas en función de los valores que adoptan las Componentes o realizar un análisis de conglomeración. Más adelante se verá el aporte de Componentes Principales al Análisis Factorial. En función de esto, será posible interpretar las nuevas variables en términos de característica subyacentes de las empresas.

2.3 Análisis factorial

El objetivo del Análisis Factorial es descubrir factores latentes u ocultos que generan la estructura de correlaciones de un conjunto de variables. El analista intenta, en la medida de lo posible, dar sentido a estos factores en términos de orientaciones o características de las empresas, conduciendo de esta forma el análisis a un proceso tipificatorio de las mismas.

Como se indicó precedentemente, el tema Análisis Factorial se presentará en las dos versiones siguientes:

1. Análisis Factorial con factores comunes y factores específicos.
2. Análisis Factorial en Componentes Principales.

El primer enfoque es el que realmente se desarrolló en el campo de la Estadística para cumplir los propósitos mencionados más arriba. No obstante, el segundo enfoque también cumple con estos objetivos, en virtud de una serie de supuestos y resultados matemáticos que hacen que el tema Componentes Principales brinde grandes posibilidades, fundamentalmente en la extracción de los factores.

A efectos de una primera aproximación al tema, en ambos enfoques se asume que los factores latentes son independientes y que actúan en forma lineal sobre las variables. En el Análisis con factores específicos puede asumirse otro tipo de relación entre factores y variables; en el segundo Análisis, solamente si se adopta el supuesto de linealidad puede hacer su aporte el tema Componentes Principales.

Se exponen a continuación ambos enfoques.

2.3.1. Análisis Factorial con Factores comunes y factores específicos

Introducción

Como ya se dijo precedentemente, este análisis fue desarrollado expresamente para detectar factores ocultos o latentes, distinguiéndose entre comunes y específicos, de un conjunto de variables. Estos factores se suponen independientes y linealmente relacionados a las variables. Es decir, todas las variables se expresan en forma lineal, en función de un conjunto pequeño de factores comunes, más un factor específico que resume la parte no común de cada variable con las restantes.

Se calcula la matriz de correlaciones entre las variables originales y los factores comunes. Cada columna de esta matriz contiene a los coeficientes de correlación entre un factor y todas las variables. Por lo tanto, cada columna identifica a un factor.

El análisis del sentido o la interpretación de los factores, se efectúa sobre esta matriz, considerando el signo y la intensidad de la correlación de cada factor con las variables originales.

Los supuestos de linealidad de la relación entre variables y factores y de independencia entre factores, permiten separar la varianza de cada variable en dos partes. La primera se denomina "comunalidad" e identifica la contribución de los factores comunes a la varianza de cada variable. La segunda parte de la varianza se denomina "especificidad" y expresa cuánto de específico conserva cada variable, lo que no es explicado por el conjunto de factores extraídos.

Cada comunalidad, a su vez, puede expresarse como suma de las contribuciones de cada factor. Con estas contribuciones se hace el análisis de las variables a los efectos de determinar por qué factores ellas resultan mejor explicadas.

Si el conjunto de factores extraído es pequeño, y si éstos explican suficientemente bien a las variables originales, se habrá ganado en simplicidad. Si además, se logra determinar el sentido de los factores en

términos de orientaciones o características de las empresas, se habrá dado un paso muy importante en el proceso de tipificación de las mismas.

No obstante, puede ocurrir que el número de factores extraído sea insuficiente. En tal caso, puede plantearse un nuevo modelo, en el que intervengan más factores, con el cual es de esperar que las variables originales resulten mejor explicadas. Dado que este modelo contempla un factor específico que actúa como receptáculo de toda aquella parte de las respuestas que no son explicadas por los factores, está "abierta" la posibilidad de que al extraer más factores este receptáculo disminuya, lográndose mayor comunalidad para las variables (y, por lo tanto, menor especificidad de las mismas). Es decir, con más factores, es posible que las variables resulten mejor explicadas.

Estas consideraciones son las que llevan a Cattell* a denotar por "modelo abierto" a esta forma de análisis para contraponerla al enfoque proporcionado por Componentes Principales, que veremos más adelante.

Conceptos Teóricos Básicos

La información básica de la cual se partirá, provendrá de la matriz X_{pxN} de observaciones. Se denotará por x_{px1} el vector de valores estandarizado correspondiente a las empresas y se asumirá que existen q factores latentes ($q < p$) que actúan en forma lineal sobre las variables respuestas. Es decir que cada variable original X_i se expresa como combinación lineal de esos factores aún desconocidos, más un factor específico, propio de cada variable:

$$X_i = \sum_{j=1}^{q} m_{ij} f_j + e_i \qquad (I)$$

$$i = \overline{1,p}$$

en donde f_j, $j = \overline{1,q}$ son los factores comunes a las respuestas, las cuales se suponen no correlacionadas y estandards. Por su parte, los términos e_i, $i = \overline{1,p}$ son factores específicos que resumen la parte no común de cada variable con las restantes (es decir: lo específico que conserva cada variable en el modelo de q factores). También estos elementos se suponen no correlacionados entre ellos ni con los factores comunes.

Finalmente, m_{ij} expresa el "peso" de cada factor sobre las variables en un sentido a ser desarrollado más adelante.

En notación matricial, el modelo se expresa de la siguiente manera:

$$x = M \cdot f + e \qquad (II)$$

en donde x_{px1} es el vector observación estandarizado, f_{qx1} es el vector de factores comunes, M_{pxq} es la matriz de "pesos" y e_{px1} es el vector de factores específicos.

De acuerdo a la expresión (I), se demuestra** que:

$$1 = \text{Varianza}\ (X_i) = \sum_{j=1}^{q} m_{ij}^2 + \psi_i \qquad (III)$$

$$i = \overline{1,p}$$

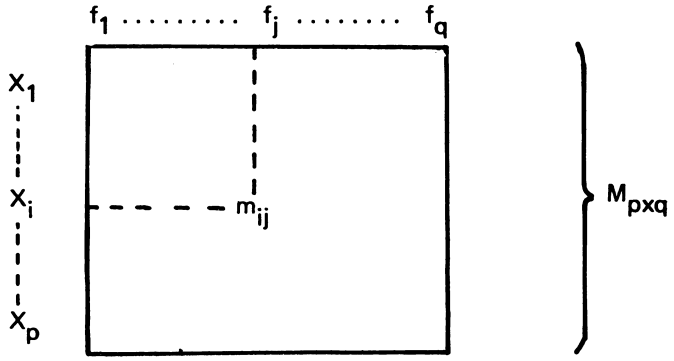
---

* Véase (2), pp. 190 - 215

** La Varianza de una suma de variables independientes es igual a la suma de las varianzas de cada una de ellas. Véase (12), pág. 83.

donde ψ_i = Varianza (e_i) se denomina "especificidad" de la variable X_i y $\sum_{j=1}^{q} m_{ij}^2$ se denomina "comunalidad" de la variable X_i; estas comunalidades indican la aptitud que tienen los q factores comunes para explicar la dispersión de cada variable.

La matriz de correlaciones entre el vector observación x y el vector de factores f es la matriz M *. Por ende, m_{ij} es el coeficiente de correlación entre la i-ésima variable respuesta y el j-ésimo factor común:



$$\text{CORR}\ (x, f) \quad = \quad M_{pxq}$$

$$\text{corr}\ (X_i, f_j) \quad = \quad m_{ij}$$

Recordando que R es la matriz de correlaciones originales, en función de la expresión (II) se demuestra ** que:

$$R \quad = \quad M \cdot M^T \ + \ \psi \qquad\qquad (IV)$$

donde ψ es una matriz de dimensiones pxp cuyos elementos de la diagonal principal son las especificidades ψ_i y el resto son todos ceros:

$$\psi \quad = \quad \begin{bmatrix} \psi_1 & & 0 \\ & \ddots & \\ 0 & & \psi_p \end{bmatrix}$$

Por lo tanto, si r_{ij} es el coeficiente de correlación entre las respuestas x_i y x_j, de la expresión (IV) resulta:

$$r_{ij} \quad = \quad \sum_{k=1}^{q} m_{ik}\ m_{jk} \qquad \text{para } i \neq j \qquad (V)$$

---

* Véase (10), pág. 262.

** IBIDEM, pág. 262.

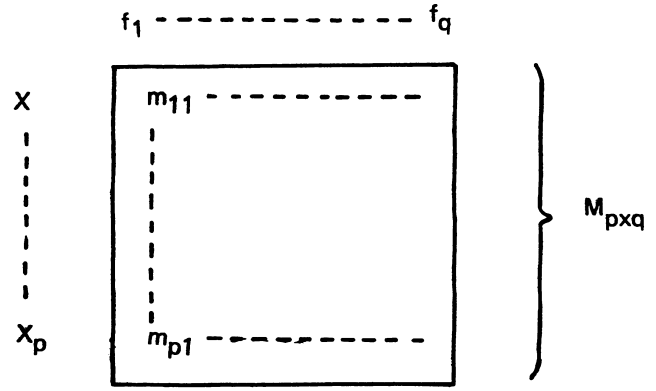
Si $i = j$, la expresión (V) se convierte en la expresión (III), dado que los elementos de la diagonal principal de R son iguales a 1.

De modo que, el producto matricial $M \cdot M^T$ reestablece a la matriz R de correlaciones originales, salvo los elementos de la diagonal principal en la cual la varianza de cada variable se reestablece con la suma de la comunalidad más la especificidad de cada variable. De otra forma, la expresión (V) indica que mediante las correlaciones entre variables y factores m_{ij} se reestablecen las correlaciones entre variables originales.

Analizando nuevamente la expresión (III), m_{ij}^2 se interpreta como la cantidad de varianza que explica el factor j-ésimo de la variable i-ésima, ψ_i es la parte que queda sin explicar de la variable X_i por los q factores.

Adviértase que sobre un fenómeno actúan innumerables factores y que el analista en general está dispuesto a que el modelo extraiga sólo unos pocos. Por ello debe aceptarse, y por lo tanto debe dejarse cierta libertad, de que las variables conserven algo de específicas frente a estos factores. Claro está que cuanto menos específicas permanezcan las variables, mejor explicadas estarán por el modelo.

Los resultados expuestos hasta este momento se utilizan para efectuar el análisis de los factores. La matriz M proporciona información respecto de cómo se correlacionan los factores y las variables. Cada columna de M es un vector de correlaciones entre un factor y todas las variables originales. Por lo tanto cada columna identifica a un factor:



Análisis Vertical

Una vez obtenidos los resultados básicos, el analista puede efectuar este tipo de análisis sobre la matriz M, buscando el sentido de los factores en función del signo y de la intensidad de sus correlaciones con las variables.

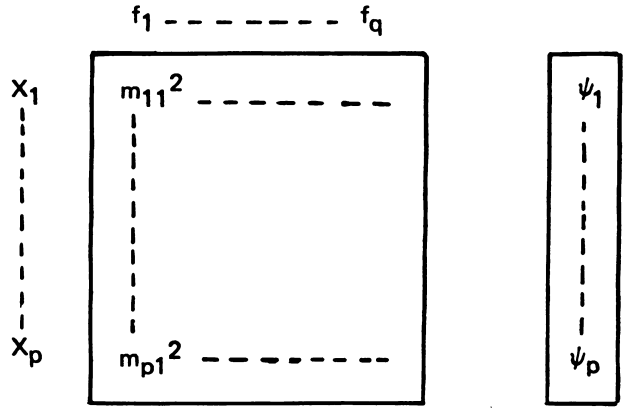
Ejemplo 1

Factores:	f_1	f_2	f_3	
Variables Originales:				
Superficie Total	.9	.2	–.1	
Superficie Chacra	–.3	.7	.2	
Vacas Cría/Stock Vacuno	.4	.4	–.8	
Porcentaje de parición en vacunos				M_{6x3}
Capital Total	.3	.2	.4	
Producto Bruto	.9	–.3	–.3	
	.8	.5	–.3	

En este ejemplo (con M_{6x3}), el factor f_1 se identifica con lo que podríamos llamar dimensión de las empresas, dada su alta correlación positiva con las variables Superficie Total, Capital Total y Producto Bruto.

Análisis Horizontal

Sea $M^{(2)}$ la matriz que se obtiene elevando al cuadrado los elementos de la matriz M. Si se le agrega una columna adicional, constituída por las especificidades ψ_i, la nueva matriz permitirá efectuar el análisis de cada variable.en función de los factores:



El analista puede también optar por efectuar su análisis sobre la matriz $M^{(2)}$ orlada con una columna de especificidades. Dado que m_{ij}^2 indica la cantidad de varianza que explica el factor f_j de la variable X_i, sumando los elementos de una misma fila de $M^{(2)}$ puede determinarse la cantidad de varianza que aporta cualquier subconjunto de factores a una variable determinada. Es decir que para cada variable se trata de determinar por qué factores está mejor explicada.

Ejemplo 2

Factores	f_1	f_2	f_3	ψ
Variables Originales:				
Superficie Total	.81 +	.04 +	.01 –	.14
Superficie Chacra	.09 –	.49 +	.04 +	.38
Vacas Cría/Stock vacuno	.16 +	.16 +	.64 –	.14
Porcentaje de parición en vacunos				
Capital Total	.09 +	.04 +	.16 +	.71
Producto Bruto	.81 +	.09 –	.09 –	.01
	.64 +	.25 –	.09 –	.02

El ejemplo, continuación del presentado anteriormente, trata de ejemplificar el análisis de la variable Superficie dedicada a Chacra. Evidentemente, está poco explicada por el factor dimensión (9%) y aparece como muy explicada por el segundo factor (49%). También se observa que ella permanece bastante específica en este modelo de tres factores. El signo colocado a la derecha de cada elemento de $M^{(2)}$ es el signo de la correlación entre el factor y la variable.

Obsérvese que con un poco de cuidado, ambos análisis Horizontal y Vertical pueden realizarse sobre la matriz $M^{(2)}$. Para ello se recomienda* colocar al costado derecho de cada valor m_{ij}^2 el signo del correspondiente valor m_{ij}. De esta manera, dado que los números m_{ij}^2 son positivos, no se pierde la información de la "dirección" en que se correlacionan las variables y los factores.

No obstante estas consideraciones, debe enfatizarse que la base de este análisis está en la matriz M de pesos o correlaciones. Uno de los objetivos primordiales del analista debe ser obtenerla y con sus columnas identificar a los factores.

Este problema de cálculo es central y a él puede contribuir un enfoque que parta de Componentes Principales. Eso se discutirá más adelante.

Se expondrá a continuación un resumen del método de Lawley, con el cual se obtienen estimadores para los elementos de las matrices M y ψ. Se basa en el método de máxima verosimilitud para obtener estimadores, para lo cual debe resolverse un sistema de ecuaciones. Rao y Maxwell propusieron un proceso iterativo para resolver tal sistema, el cual consiste en resolver primero el modelo con un solo factor; luego, en función de esta solución se resuelve el modelo con dos factores, y así sucesivamente, hasta resolver el modelo con q factores. Cada resolución o etapa de este proceso también se efectúa en forma iterativa, basándose en la extracción de Componentes Principales.

Las soluciones de los modelos cambian al variar el número de factores que se extraen. Dado que los factores están caracterizados por las columnas de M y los elementos de éstas cambian numéricamente, suele ocurrir que el primer factor de un modelo con q factores difiere del primer factor de un modelo con q + 1 factores. El siguiente ejemplo numérico extraído del libro de Morrison (10), página 273, ejemplifica la la situación expuesta anteriormente:

---

* Véase (3), pág. 163

Variables	Modelo con un factor	Modelo con dos factores	
	Factor	Factor 1	Factor 2
X_1	.637	.674	.158
X_2	.584	.715	.650
X_3	.959	.947	–.157
X_4	.960	.939	–.217
X_5	.929	.908	–.189
X_6	.934	.920	–.159

Dado que los estimadores de M y ψ se obtienen por el método de máxima verosimilitud, la capacidad del modelo (cualquiera sea el número de factores extraídos), de reconstruir la matriz R de correlaciones puede docimarse mediante una prueba chi-cuadrado en la cual se comparan los determinantes de las matrices R y $\hat{M}\,\hat{M}^T + \hat{\psi}$ (donde $\hat{M}$ significa "estimador de máxima verosimilitud de M")*. Con esta prueba, se testea la suficiencia del modelo con el número de factores extraído. En caso de rechazo puede continuarse el proceso extrayendo más factores y volviendo a docimar. Estas dócimas consecutivas, utilizando la misma información básica, trae aparejado aumento en los niveles de significación. Deberían usarse niveles muy bajos en caso de efectuarse muchas dócimas consecutivas o docimar solamente el modelo con un número preestablecido de factores.

Varias consideraciones pueden ayudar a delimitar el número de factores a prefijar por el analista. El análisis y la interpretación de los factores pueden sugerir el descarte de algunos. El grado de especificidad de algunas variables puede sugerir, por lo contrario, la insuficiencia de los factores extraídos para explicar el conjunto de respuestas. En la elección de los atributos o variables investigados está implícito el conjunto de factores relevantes. Cuando se decide incluir cierto conjunto de variables, se está pensando que ellas contemplan algunos aspectos de las empresas que se creen importantes. Luego el análisis, junto a las dócimas mencionadas más arriba, confirmará o no las ideas e hipótesis previas del analista.

Si para cada empresa analizada se evalúan los factores comunes, puede pensarse en efectuar un Análisis de Conglomeración. Para la evaluación de los factores se dispone de dos métodos** que no son exactos y que proporcionan soluciones aproximadamente iguales sólo si se cumplen algunas condiciones establecidas con las matrices M y ψ.

Resumen

El modelo con factores comunes y factores específicos se resuelve determinando a priori el número de factores que el analista está dispuesto a extraer. Los factores específicos e_i, así como las especificidades ψ_i, contemplan aquella parte de las respuestas no explicadas por los factores comunes y están "abiertas" a nuevas resoluciones del modelo en el caso de que el ajuste no sea bueno. Por ello se ha llamado a este modelo "modelo abierto". Así se lo contrapone al análisis en Componentes Principales o "modelo cerrado", ya que para éste el número de factores está determinado por el número de variables.

* Véase (10), 268
** Véase (10), 291

La matriz M de correlaciones entre variables y factores proporciona la información necesaria para determinar el sentido de los factores, en términos de orientaciones o caracterísitcas subyacentes de las empresas, así como el análisis individual de las variables respuestas.

Cada uno de los factores extraído puede evaluarse y con ellos realizar un Análisis de Conglomeración entre empresas.

El resumen de la información a unos pocos factores, el análisis de éstos y la conglomeración de las empresas pueden concebirse como etapas importantes de una tipificación objetiva.

2.3.2 Análisis Factorial en Componentes Principales

Introducción

Los objetivos de este enfoque del Análisis Factorial son los mismos que los del enfoque presentado anteriormente. Es decir, descubrir factores latentes cuyo análisis contribuya a un proceso tipificatorio de empresas.

Las posibilidades surgen de algunos resultados matemáticos vinculados al tema Componentes Principales, en función de los cuales se logra definir un número de factores latentes igual al número de variables originales. Por lo tanto, debe enfatizarse que, dado que se fundamenta en un tema que por su naturaleza misma es rígido (rotación de ejes) y que no se desarrolló para cumplir con los objetivos del Análisis Factorial, deben agregarse una serie de supuestos para que Componentes Principales brinde posibilidades tanto en el plano conceptual como en la extracción y análisis de los factores.

Los supuestos sobre los cuales se fundamenta el aporte de Componentes Principales al Análisis Factorial son: que los factores latentes actúan en forma lineal sobre las variables, que son factores independientes, que su número es igual al número de variables originales y que son los únicos que actúan sobre las variables. Es decir, cada variable respuesta se expresa en forma lineal en función de un conjunto de factores comunes independientes y estos son los únicos que actúan sobre las variables. En comparación al tema recién expuesto, cabe destacar: la linealidad es un supuesto esencial, se generan tantos factores como variables originales hay, no hay especificidad.

Claramente, esta es una concepción rígida del Análisis Factorial. Dado que sobre un fenómeno actúan innumerables factores, se debe restringir el número de los mismos sin la posibilidad de disponer de un receptáculo que resuma a los restantes factores no considerados.

Todos los factores que actúan sobre las variables se extraen de una vez de la matriz R de correlaciones entre variables originales. Para ello se calcula una matriz M de correlaciones entre las respuestas y los factores comunes, en función de los vectores y valores propios de la matriz R. Es decir, la obtención de la matriz M comienza por la extracción de las Componentes Principales.

Esta matriz M es cuadrada de dimensiones pxp y cada una de sus columnas identifica a un factor. El análisis del sentido o la interpretación de los factores se efectúa sobre la matriz M en función del signo y de la intensidad de la correlación de cada factor con las variables originales.

Los supuestos de linealidad de la relación entre variables y factores y de independencia entre factores permiten expresar la varianza de cada variable como suma de aportes independientes de cada uno de los factores. Con estas contribuciones se hace el análisis individual de las variables a los efectos de determinar cuáles factores explican mejor a las mismas.

Ambos análisis: el de los factores y el de las variables individuales, contribuyen a delimitar el número de factores a retener por el analista. Aunque no se trata del objetivo principal del Análisis Factorial, en general también con su uso se trata de simplificar, reduciendo las p variables originales a un conjunto menor

de factores comunes. Estos, habitualmente, resultan ser los primeros. Es decir, los que corresponden a las Componentes Principales que explican mayor porcentaje de la varianza total. Los últimos factores, dado que sus correlaciones con las variables pierden intensidad, resultan difíciles de interpretar y sus contribuciones a las varianzas de las respuestas son de poca importancia.

Los factores comunes son fácilmente evaluables para cada empresa, pudiéndose, por lo tanto, efectuar luego un Análisis de Conglomeración. Una primera idea sobre la construcción de los grupos puede darla un gráfico en donde las empresas se representan en el plano de los dos primeros factores*.

Conceptos teóricos básicos

Al tratar el tema Componentes Principales se indicó que éstas son las coordenadas transformadas, a través de una matriz ortogonal, de las coordenadas o valores originales correspondientes a las empresas:

$$y = A^T x \qquad (I)$$

Dado que A^T es una matriz ortogonal, su inversa es la matriz A (cuyas columnas son las filas de A^T). Por lo tanto, la expresión (I) se escribe también de esta otra forma:

$$Ay = A\ A^T x$$

$$Ay = I_{pxp} x$$

es decir:

$$x = Ay \qquad (II)$$

donde I_{pxp} es la matriz identidad de dimensiones pxp.

El vector y, cuyas coordenadas son $Y_1, Y_2, \ldots . Y_p$ ("Componentes Principales"), tiene como valor esperado al vector **"cero"** (todas sus coordenadas son ceros) y matriz de varianzas y covarianzas:

$$L = \begin{bmatrix} l_1 & & 0 \\ & \ddots & \\ 0 & & l_p \end{bmatrix}$$

dado que varianza $(Y_i) = l_i$ (i-ésimo valor propio de R) y corr $(Y_i, Y_j) = 0$ si $i \neq j$.

---

* Véase Kaminsky M. en (7), Vol. 2, pp. 57, 76, 80 y 145.

Definiendo las matrices $L^{1/2}$ y $L^{-1/2}$ de la siguiente manera:

$$L^{1/2} = \begin{bmatrix} \sqrt{l_i} & & 0 \\ & \ddots & \\ 0 & & \sqrt{l_p} \end{bmatrix} \qquad L^{-1/2} = \begin{bmatrix} \frac{1}{\sqrt{l_1}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{l_p}} \end{bmatrix}$$

puede verse que $L^{1/2} \cdot L^{-1/2} = I_{pxp}$ y que el vector $f = L^{-1/2}$ y está estandarizado*.

En función de estos resultados, la expresión (II) puede ponerse de esta otra forma:

$$x = A \frac{L^{1/2}}{I} \frac{L^{-1/2}}{} \cdot y$$

$$x = A \cdot L^{1/2} \cdot f \qquad \text{(III)}$$

Debe recordarse que las columnas de A son los sucesivos vectores propios normalizados de R, en orden de importancia, de aucerdo a las varianzas que explican (valores propios). Puede ahora definirse la matriz:

$$M = A \cdot L.^{1/2}$$

donde la i-ésima columna de M es igual a la i-ésima columna de A, (o sea, el i-ésimo vector propio normalizado de R), multiplicado por la raíz cuadrada del i-ésimo valor propio de R. Entonces la ecuación (III) se escribe:

$$X = M \cdot f \qquad \text{(IV)}$$

donde x_{px1} es el vector observación estandarizado, f_{px1} es un vector de p variables estandards no correlacionadas que en adelante llamaremos **"factores"** y M_{pxp} es la matriz de "pesos" de los factores sobre las variables originales.

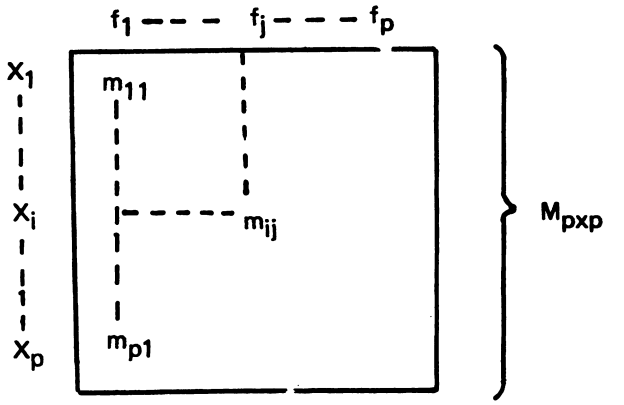
La expresión matricial (IV) representa al modelo en Componentes Principales que define a los factores latentes. Como en el Análisis Factorial convencional, M es la matriz de correlaciones entre las p variables y los p factores.

---

* En efecto, $f_i = \frac{Y_i}{\sqrt{l_i}}$, luego $E(f_i) = 0$, $Var(f_i) = \frac{1}{l_i} Var(Y_i) = \frac{1}{l_i} \cdot l_i = 1$ y matriz $COV(f) = I_{pxp}$.

$$\text{CORR}\ (x, f) = M$$

$$\text{corr}\ (X_i, f_j) = m_{ij}$$



La expresión (IV) también se interpreta de esta manera: "Cada variable original es función lineal de los p factores"

$$X_i = \sum_{j=1}^{p} m_{ij}\, f_j, \quad i = \overline{1,p} \qquad \text{y por lo tanto:}$$

$$1 = \text{Var}\ (X_i) = \sum_{j=1}^{p} m_{ij}^2 \qquad i = \overline{1,p} \qquad \text{(V)}$$

dado que Var $(f_i) = 1$ y cov $(f_i, f_j) = 0$.

Sea R la matriz de correlaciones de las variables respuestas: entonces, en función de (IV), se tiene* que:

$$R = MM^T \qquad \text{(VI)}$$

Por lo tanto, si r_{ij} es el coeficiente de correlación entre las variables X_i y X_j, la expresión (VI) indica que

$$r_{ij} = \sum_{k=1}^{p} m_{ik}\, m_{jk} \qquad \text{(VII)}$$

para todo $i = \overline{1,p}$ y $j = \overline{1,p}$. Si $i = j$, retomamos (V).

---

* Véase (10), pág. 227

De modo que, el producto matricial MM^T reestablece completamente la matriz R de correlaciones originales incluyendo la diagonal principal. De otro modo, en función de las correlaciones entre variables y factores m_{ij} se reestablecen las correlaciones entre variables originales.

La expresión (V) establece que m_{ij}^2 es la parte de la varianza de X_i que es explicada por el factor f_j. La suma de los aportes de los p factores, explica completamente la varianza de cada variable.

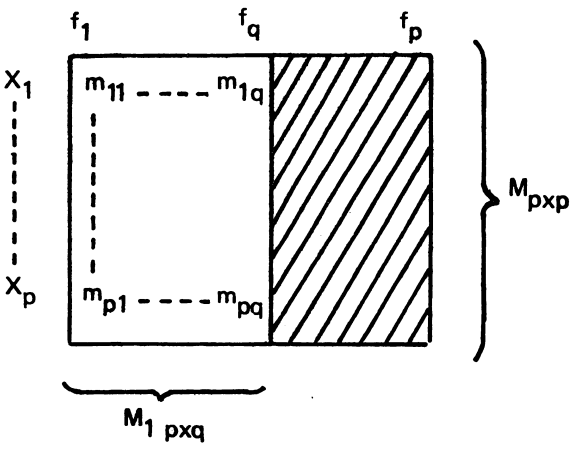
Los resultados expuestos hasta este momento implican que el modelo encontrado por aplicación de Componentes Principales se resuelve por estrictamente p factores (igual número que variables). La rigidez del método resulta del supuesto de un número de factores estricto y de que en función de todos ellos se restituyen las correlaciones y varianzas originales. Luego todos los factores deberían entrar en el análisis, que se realizaría en función de la matriz M y $M^{(2)}$ definida como en el caso 1. Se busca de esta manera el sentido y la interpretación de los factores mediante las correlaciones m_{ij} y el aporte a las varianzas de las variables. Es decir: se buscaría hacer tanto el análisis horizontal como el vertical con M y $M^{(2)}$, igual que en el enfoque anterior, pero ahora con mayor número de factores.

Sin embargo, el número de factores que se conservan para trabajos tipificatorios en general es bastante inferior al número de variables. Igual que en el Análisis Factorial convencional, el número de factores que se retienen se delimita en función de varias consideraciones. Dado que cada factor proviene de una Componente Principal, una primera idea la proporciona el porcentaje explicativo de los primeros factores. Podría decidirse, por ejemplo, conservar los primeros q factores que explican el 85% de la varianza total. Además, en la práctica, sólo los primeros factores son susceptibles de interpretación clara en términos de orientaciones de las empresas. Los últimos factores normalmente se vuelven difíciles de intrepretar, dado que sus correlaciones con las variables pierden intensidad. Por la misma razón, sus contribuciones a las varianzas de las variables se vuelven irrelevantes.

Se construyen entonces las matrices M y $M^{(2)}$, pero ahora sólo con las columnas correspondientes a los factores retenidos. En la matriz $M^{(2)}$ se suman las filas colocando estos resultados como una columna más, a los efectos de visualizar el aporte de los factores retenidos a las varianzas de las variables.

De modo que, el análisis sugiere conservar algunos factores. Eventualmente puede tomarse la decisión de conservar alguno más de los restantes, marcándose así una diferencia sustancial con el Análisis Factorial convencional, en el cual más factores implican cambiar en alguna medida los factores ya extraídos.

Al mantener unos pocos factores, la matriz M pierde columnas y su capacidad para reconstruir R se ve disminuída.



$$R \doteq M_1 M_1^T$$

Un número adecuado de factores retenidos debería proporcionar un producto matricial $M_1.M_1^T$ suficientemente próximo a la matriz R.

No se dispone* en este enfoque, de una prueba de hipótesis que docime la bondad de ajuste del modelo que resulta de conservar los primeros q factores:

$$X_i = \sum_{j=1}^{q} m_{ij} f_j \qquad i = \overline{1,p} \qquad q < p$$

Sólo mediante consideraciones prácticas del tipo de las ya presentadas, se logrará decidir esta cuestión.

A diferencia con el Análisis Factorial convencional, la evaluación de los factores se realiza en forma exacta, difiriendo apenas por una constante de la evaluación de las correspondientes Componentes Principales. Al exponer Componentes Principales, quedó indicado que éstas se obtienen mediante una transformación de coordenadas a través de una matriz ortogonal A^T cuyas filas son los vectores propios normalizados de R:

$$y = A^T x$$

A los efectos de hacer aparecer los factores, se sugirió otro cambio de variable, a saber:

$$f = L^{-1/2} y$$

Por lo tanto, el factor i-ésimo se evalúa dividiendo la Componente Principal i-ésima Y_i por la raíz cuadrada del i-ésimo valor propio l_i de R:

$$f_i = \frac{Y_i}{\sqrt{l_i}} \qquad i = \overline{1,p}$$

Una vez evaluados los factores para esta empresa, puede efectuarse un Análisis de Conglomeración como una etapa más en el proceso de tipificación.

Resumen

El enfoque de Análisis Factorial en términos de Componentes Principales, ha sido llamado por Cattell** "modelo cerrado" en virtud de la rigidez de sus supuestos. En primer lugar, se asume que el número de factores latentes es igual al número de variables originales; en segundo lugar, estos factores son los únicos que actúan sobre las empresas, no estando contemplada la posibilidad de que existan más factores mediante algún término incluído en el modelo.

---

* Algunas pruebas en situaciones particulares pueden verse en (10), pp. 251 - 254

** Véase (2), op. cit., pág. 198.

Dada la matriz de observaciones, se calcula R y el modelo se resuelve de una vez (es decir, se extraen todos los factores) calculando la matriz M cuyas columnas son los vectores propios de R, multiplicados por la raíz cuadrada de los correspondientes valores propios. Estas columnas identifican a los p factores, los cuales tienen la capacidad de explicar completamente a las variables respuestas, así como de reconstruir las correlaciones originales.

El porcentaje explicativo de la varianza total brindado por los primeros factores, el análisis e interpretación de los mismos y el análisis de las variables individuales, permite delimitar un número aceptable de factores, a retener por el analista.

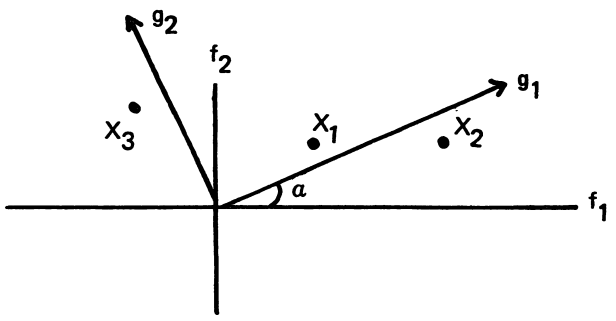
Estos factores retenidos para fines tipificatorios, son fácilmente evaluables, en función de las correspondientes Componentes Principales. Luego puede realizarse con ellos un Análisis de Conglomeración si el número de variables originales es muy numeroso

2.3.3. Rotación de los Factores

Introducción

En los dos enfoques presentados del Análisis Factorial, la extracción de los factores es un **problema indeterminado.** En efecto, encontrada una solución, o sea un conjunto de q factores que satisfaga las condiciones de cualquiera de los modelos presentados, puede obtenerse otra solución sólo con multiplicar el vector de q factores por una matriz ortogonal H. La matriz M, cuyas columnas identifican a los factores, resulta a su vez multiplicada por la matriz traspuesta de H, cambiando de esta forma las correlaciones entre factores y variables. Si en el espacio de los factores (q dimensiones) interpretamos las correlaciones de las variables con éstos como coordenadas, multiplicar los q factores por una matriz ortogonal significa hacer una rotación o giro de los ejes (factores). Cada variable pasa a tener nuevas coordenadas, o sea, nuevas correlaciones con los factores.

Mediante algunos criterios podrían intentarse rotaciones de los factores que permitan su mejor interpretación. Por ejemplo, algunas variables resultarían aún más correlacionadas con un factor y mucho menos con otros, etc. En el siguiente gráfico se intenta ejemplificar una rotación de factores a una estructura más simple de interpretar.



	f_1	f_2	
X_1	.4	.4	
X_2	.8	.4	M
X_3	−.4	.5	

Sean f_1 y f_2 los dos factores retenidos inicialmente. Sean las variables X_1, X_2 y X_3 cada una con sus correspondientes correlaciones con f_1 y f_2. Graficando estas correlaciones como coordenadas, de cada variable se obtiene un punto en el plano de f_1 y f_2.

La variable X_2 tiene correlación alta con el factor f_1 y las tres variables tienen correlaciones mediana con el factor f_2.

Sean g_1 y g_2 los factores girados un ángulo α. Las variables X_1 y X_2 pasan a tener ambas una correlación importante con g_1 y prácticamente nada de correlación con g_2. En cambio, X_3 resulta tener gran correlación con g_2 y muy poca correlación con g_1.

En estas condiciones, en el análisis del factor g_1, las variables X_1 y X_2 tendrían un peso fundamental, mientras que la variable X_3 aportaría muy poco. Para el análisis de g_2, la variable X_3 sería importante y las variables X_1 y X_2 serían irrelevantes.

Un giro de 30 grados se representa por la matriz:

$$H = \begin{bmatrix} .9 & .5 \\ -.5 & .9 \end{bmatrix}$$

en donde las cifras exactas han sido redondeadas a un decimal.

Las nuevas correlaciones se obtienen multiplicando la matriz M por la traspuesta de H:

$$\begin{bmatrix} .4 & .4 \\ .8 & .4 \\ -.4 & .5 \end{bmatrix} \begin{bmatrix} .9 & -.5 \\ .5 & .9 \end{bmatrix} = \begin{bmatrix} .56 & .16 \\ .92 & -.04 \\ -.11 & .75 \end{bmatrix}$$

Es decir:

	g_1	g_2
X_1	.56	.16
X_2	.92	–.04
X_3	–.11	.75

Este giro de 30 grados es suficiente para una mejor interpretación de los factores.

En este ejemplo, el giro de los ejes fue sugerido por el gráfico. En caso de retener más factores, los gráficos serían imposibles de realizar. Podrían representarse todas las combinaciones de dos factores posibles y con ellos efectuar el estudio de los mejores giros. Las complicaciones de este artificio no lo hacen recomendable. Piénsese que con 10 factores retenidos deberían construirse 45 gráficos.

Una técnica analítica para obtener una rotación de los factores que conduzca a una estructura más simple de analizar, es la que se conoce como "Varimax". Consiste en efectuar rotaciones de dos factores cada vez, hasta lograr un valor máximo para la suma de las varianzas de las columnas de $M^{(2)}$ en la cual se ha ponderado cada elemento por la correspondiente comunalidad. Este es el mecanismo de rotación que se emplea en el Programa de cómputo disponible en este momento en IICA, Uruguay.

Conceptos teóricos básicos

En cualquiera de los enfoques presentados, la extracción de los factores es un problema indeterminado. En efecto, multiplicando el vector formado por los q factores retenidos por una matriz ortogonal H se obtiene otra solución. A continuación, será analizada esta propiedad para cada uno de los modelos.

i) **Modelo abierto**

Recuérdese que el modelo abierto se expresa, en notación matricial, de la siguiente manera:

$$x = M.f + e \qquad (I)$$

Sea H una matriz ortogonal de dimensiones qxq. Entonces, si f es un vector estandard, también lo es $g = H.f$*. Por lo tanto, se puede escribir la expresión (I) de esta otra forma:

$$x = MH^T.H.f + e$$

dado que:

$$H^T.H = I_{qxq}$$

Sustituyendo $g = H.f$ y $M_1 = MH^T$ en la expresión anterior, se obtiene:

$$x = M_1 g + e$$

con g un vector estandard y M_1 conservando la capacidad de reconstruir R. En efecto,

$$\begin{aligned} R &= MM^T + \psi \\ &= MH^T HM^T + \psi \\ &= M_1 \times M_1{}^T + \psi \end{aligned}$$

El vector g es la nueva solución del modelo.

ii) **Modelo cerrado**

En notación matricial el modelo cerrado se expresa por:

$$x = M.f \qquad (II)$$

donde la matriz M tiene dimensiones pxp, es decir, suponemos que se retienen todos los factores.

Sea la matriz H ortogonal de dimensiones pxp. Luego el vector $g = H.f$ es un vector estandard.

---

* Véase (5), pág. 69

La expresión (II) puede escribirse de esta forma:

$$x = MH^T H.f, \text{ donde } H^T H = I_{pxp}$$

Si denotamos por M_1 el producto matricial $M.H^T$, entonces la expresión (II) se escribe también en esta forma:

$$x = M_1 g$$

donde g es un vector estandard y M_1 tiene la capacidad de reconstruir a la matriz R. En efecto:

$$\begin{aligned} R &= MM^T \\ &= MH^T HM^T \\ &= MH^T . (MH^T)^T \\ &= M_1 . M_1^T \end{aligned}$$

El vector g es la nueva solución del modelo.

Si se retienen sólo los q primeros factores, la matriz H puede definirse en forma particionada:

$$H_{pxp} = \left[\begin{array}{c|c} K_{qxq} & 0 \\ \hline 0 & I \end{array}\right]$$

donde Kqxq es una matriz ortogonal en un espacio q-dimensional y la matriz identidad I tiene dimensiones $(p - q) \times (p - q)$. Esta matriz H efectúa la rotación de los primeros q factores dejando los restantes inmóviles.

De modo que los factores no están unívocamente determinados. Dado un conjunto de factores, solución de cualquiera de los modelos, se obtiene otro conjunto de factores, que también es solución, por aplicación de una matriz ortogonal.

Los nuevos factores se identifican ahora con las columnas de la matriz $M_1 = MH^T$, o sea, la matriz de factores original, multiplicada por la matriz traspuesta de la que proporciona la rotación de los factores. El análisis se efectúa sobre la nueva matriz M_1.

Si se identifica cada uno de los q factores por los ejes ortogonales de un espacio de q dimensiones, cada variable se representa por un punto cuyas coordenadas son sus correlaciones, con cada uno de los q factores. Una matriz H ortogonal en este espacio, se interpreta como un giro de ejes. O sea, un giro de los factores, en donde los puntos que representan a las variables permanecen fijos, y los ejes cambian de posición manteniendo la perpendicularidad entre ellos. En estos nuevos ejes las variables tienen otras coordenadas, o sea, otros coeficientes de correlación con los nuevos factores, dados por las columnas de la nueva matriz M_1.

Se continúa de esta manera hasta girar por parejas los q factores. Es decir, se realizan $\frac{1}{2}$ q (q − 1) giros, consistiendo cada uno de ellos en el giro de dos factores, dejando los restantes inmóviles. En este momento del proceso, se dice que se ha completado un ciclo, obteniéndose un valor final del ciclo para v. Se repiten los ciclos hasta que todos los ángulos de los giros planos dentro de un ciclo sean inferiores a una cierta cantidad prefijada.

Kaiser demostró que el valor v no puede exceder el valor $\frac{q-1}{q}$

Por lo tanto, como cada rotación plana aumenta el valor de v o lo deja igual (en cada rotación se busca el máximo para v), este método iterativo asegura la convergencia del valor v. En la práctica se calcula el valor de v después de cada ciclo y se detiene el proceso después de obtener cuatro valores de v que no difieran en más de una cantidad prefijada.

El análisis se realiza con la matriz de factores rotada.

2.3.4. Aplicación del Análisis Factorial en Componentes Principales

Cuando se trató el tema "Componentes Principales" se hizo referencia a un programa, desarrollado por IBM, el cual extrae, de la matriz de correlaciones entre variables originales, los vectores y valores propios que identifican a las nuevas variables. En realidad el programa realiza el Análisis Factorial en el sentido de que, además de extraer las Componentes Principales, calcula la matriz de correlaciones M entre variables respuestas y factores, y realiza la rotación de los factores con el criterio "Varimax" expuesto precedentemente.

La siguiente aplicación presenta la salida impresa de este programa, a los efectos de ilustrar al usuario. Se trata de un ejemplo numérico en el cual se ha intentado exponer situaciones que no estén fuera de la realidad.

En esta presentación se asume que se relevan las siguientes cinco variables, para cada una de siete empresas:

X_1 = Superficie total

X_2 = Producto Bruto

X_3 = Capital Total

X_4 = Porcentaje de terneros señalados

X_5 = Porcentaje de parición en vacunos

Los datos obtenidos se presentan en la siguiente matriz de observaciones:

		EMPRESAS						
		1	2	3	4	5	6	7
variables	1	500	100	300	400	600	50	200
	2	40	20	30	45	52	10	30
	3	75	45	58	91	110	18	62
	4	.70	.75	.80	.70	.65	.69	.60
	5	.81	.85	.89	.82	.78	.80	.70

Se expondrá a continuación la salida impresa del Programa, transcribiendo títulos y explicando su significado.

FACTOR ANALYSIS ..... (Nombre Identificatorio)

Se imprimen el número de casos o empresas analizadas y el número de variables investigadas en cada una.

MEANS

Se imprimen, por filas, los promedios de las variables. En caso de, por ejemplo, 25 variables, se imprimen dos filas de diez promedios y una de cinco.

STANDARD DEVIATIONS

De la misma forma que en el título anterior, se imprimen los desvíos típicos de las variables originales.

CORRELATION COEFFICIENTS

Se calcula la matriz de correlaciones de las variables respuestas y se imprime por filas. En caso de muchas variables, en cada fila se imprimen diez elementos hasta agotar los elementos de la fila de la matriz de correlaciones.

EIGENVALUES

Se imprimen los valores propios, de la matriz de correlaciones, mayores que un valor prefijado por el usuario, de mayor a menor. En este caso, el valor límite utilizado fue 1.8. Por ello se imprimieron sólo los dos valores propios mayores.

CUMULATIVE PERCENTAGE OF EIGENVALUES

Cada valor propio se divide por el número de variables originales. Estos resultados se imprimen acumulados (cada uno con todos los anteriores). Obsérvese que los dos primeros factores conservan el 98% de la varianza total.

EIGENVECTORS

Bajo los títulos VECTOR 1, etc., se imprimen por filas los vectores propios, de la matriz de correlaciones. Cada vector se corresponde con uno de los valores propios impresos antes y en el mismo orden.

FACTOR MATRIZ (k FACTORS)

Se imprime la matriz de factores. Cada columna está constituída por las correlaciones de un factor con todas las variables. Cada fila está constituída por las correlaciones entre una variable y los factores retenidos.

FACTOR ANALYSIS ..... PRUEBA

NO. OF CASES 7
NO. OF VARIABLES 5

MEANS

	307.14282	32.42856	65.57143	0.69857	0.80714

STANDARD DEVIATIONS

	204.99706	14.53567	30.18196	0.06466	0.05936

CORRELATION COEFFICIENTS

ROW 1					
	1.00000	0.94966	0.91914	–0.15627	–0.03228
ROW 2					
	0.94966	1.00000	0.99164	–0.24395	–0.12776

ROW 3
0.91914 0.99164 1.00000 –0.27109 –0.15243

ROW 4
–0.15627 –0.24395 –0.27109 1.00000 0.98009

ROW 5
–0.03228 –0.12776 –0.15243 0.98009 1.00000

EIGENVALUES
3.06447 1.83833

CUMULATIVE PERCENTAGE OF EIGENVALUES
0.61289 0.98056

EIGENVECTORS

VECTOR 1
0.52184 0.54996 0.54896 –0.27864 –0.21501

VECTOR 2
0.24969 0.18862 0.16577 0.64104 0.68093

FACTOR MATRIX (2 FACTORS)

VARIABLE 1
0.91351 0.33854

VARIABLE 2
0.96274 0.25573

VARIABLE 3
0.96099 0.22475

VARIABLE 4
–0.48778 0.86915

VARIABLE 5
–0.37638 0.92324

ITERATION CYCLE	VARIANCES
0	0.258364
1	0.460516
2	0.460516
3	0.460516
4	0.460516
5	0.460516

ROTATED FACTOR MATRIX (2 FACTORS)

VARIABLE 1
0.97417 –0.01003

VARIABLE 2
0.99058 –0.10495

VARIABLE 3
0.97788 –0.13326

VARIABLE	4	
	–0.14522	0.98604
VARIABLE	5	
	–0.02185	0.99678

CHECK ON COMMUNALITIES

VARIABLE	ORIGINAL	FINAL	DIFFERENCE
1	0.94911	0.94911	0.00000
2	0.99227	0.99227	0.00000
3	0.97401	0.97401	0.00000
4	0.99336	0.99336	0.00000
5	0.99404	0.99404	0.00000

Obsérvese que el análisis de estos factores podría realizarse ya sobre esta matriz. El primer factor (primer columna) se identifica con las variables Superficie Total, Producto Bruto y Capital Total, o sea con las variables indicadoras de dimensión o tamaño (del grupo de variables relevadas). El segundo factor (segunda columna) se identifica con las dos últimas variables, o sea con Porcentaje de terneros señalados y Porcentaje de Parición. Estas variables son indicadoras de eficiencia técnica en el manejo de vacunos.

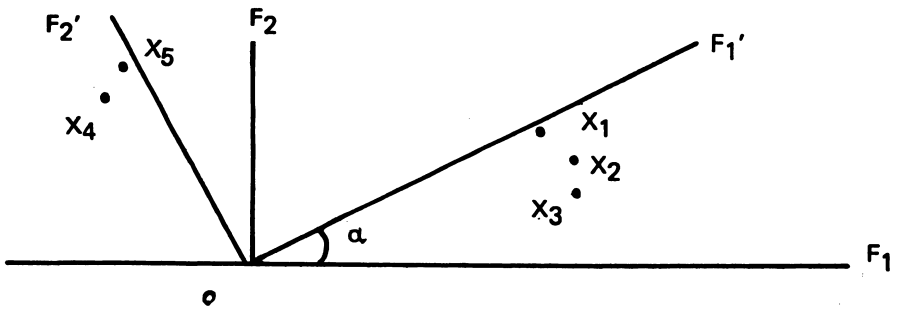
No obstante, el primer factor posee correlación negativa medianamente importante con la cuarta y quinta variable. En muchos casos esta situación podría dificultar la interpretación de los factores. Para evitar esto, el Programa continúa, efectuando la rotación de los factores a los efectos de, mediante el criterio Varimax, lograr una estructura más sencilla de analizar.

ITERATION CYCLE	VARIANCES

En dos columnas y en correspondencia aparece la cantidad de ciclos y el valor de la varianza, mencionada al exponer el criterio Varimax.

La presente salida impresa muestra que fue necesario sólo un ciclo para lograr el valor máximo para la varianza. Luego se realizaron cuatro ciclos más observándose que no hubo incremento. El programa prevé 50 ciclos.

En el siguiente gráfico se presentan las variables originales, como puntos en el espacio de los factores. Las correlaciones de cada variable con ambos factores fueron interpretados como sus coordenadas.



Se observa claramente como se agrupan por separado, las tres primeras variables por un lado, y las dos últimas por otro. El mejor giro de los factores que logre que las tres primeras variables se correlacionen más con el nuevo primer factor y menos con el segundo y que las dos últimas variables se correlacionen más con el segundo factor y menos con el primero, estará determinado por un ángulo de aproximadamente 20 grados antihorario.

ROTATED FACTOR MATRIX (k FACTORS)

De la misma forma que la matriz de factores, se imprime la nueva matriz, que se obtiene luego de efectuar la rotación de los factores.

El ejemplo presentado pretende ilustrar las posibilidades de depuración que ofrece la rotación de los factores mediante el criterio Varimax. Obsérvese que se enfatiza la interpretación de los factores al aumentar las correlaciones importantes y disminuir las correlaciones bajas. En esta nueva matriz la interpretación se hace más clara.

CHECK ON COMMUNALITIES

VARIABLE	ORIGINAL	FINAL	DIFFERENCE

Como fórma de control, se imprimen por filas el número ordinal de cada variable, su comunalidad original (la que surge de sumar por filas los elementos de la matriz de factores, elevados al cuadrado), su comunalidad final (sumando los elementos de la matriz rotada, elevados al cuadrado) y la diferencia entre ambos.

Una vez en posesión de esta salida impresa, se calculó para cada una de las siete empresas, los valores que asumen los dos factores retenidos. Estos valores se presentan en la siguiente matriz:

	EMPRESAS						
	1	2	3	4	5	6	7
FACTOR 1	.53	−1.00	−.56	.64	1.49	−1.32	.22
FACTOR 2	.32	.35	1.38	.43	.05	−.76	−1.76

Como ejemplo se calculará el valor que adopta el factor 1 para la empresa 4. Para ello, debe recordarse que el análisis se inició calculando, a partir de la matriz de observaciones, la matriz de correlaciones.

Por lo expuesto en este trabajo, el análisis debería iniciarse con el cálculo de la matriz de covarianzas de las variables. Pero como fue indicado precedentemente, si se estandarizan las variables la nueva matriz de covarianzas es la matriz de correlaciones de las variables con sus valores originales.

De modo que debe asumirse que el análisis se inició realmente con la estandarización mencionada y la construcción de una nueva matriz de valores estandarizados, todo lo cual no es necesario realizar hasta que se desee conectar los resultados con las empresas mismas.

Dado que, en función de los resultados obtenidos a partir de la matriz de correlaciones, se desea calcular al factor 1 para la empresa 4, se debe comenzar por estandarizar dicha empresa.

Empresa 4	$\bar{X}_i$	S_i	Valores Estandard Z_i
X_1 = 400	307.14282	204.99706	.45297
X_2 = 45	32.42856	14.53567	.86487
X_3 = 91	65.57143	30.18196	.84251
X_4 = .70	.69857	.06466	.02212
X_5 = .82	.80714	.05936	.21664

En la página 25 se expresó que "para evaluar el factor i-esimo se divide la componente principal i-esima por la raíz cuadrada del i-esimo valor propio l_i de R".

De modo que se necesitan l_1 y Y_1, primer valor propio y primer componente principal respectivamente.

En la página 12 se expresó "la primera fila de A^T (primer vector propio de R) "multiplicada"* por el vector de respuestas de una empresa dada proporciona Y_1 para esa empresa".

El primer vector propio de R se obtiene de la salida impresa: Eigenvectors.

Vector 1

.52184 .54996 .54896 –.27864 –.21501

El cual "multiplicado"* por el vector estandarizado de la empresa 4:

.45297 .86487 .84251 .02212 .21664

proporciona:

$$Y_1 = 1.12178$$

Como l_1 = 3.06447 (primer eigenvalue de la salida impresa)

Resulta:

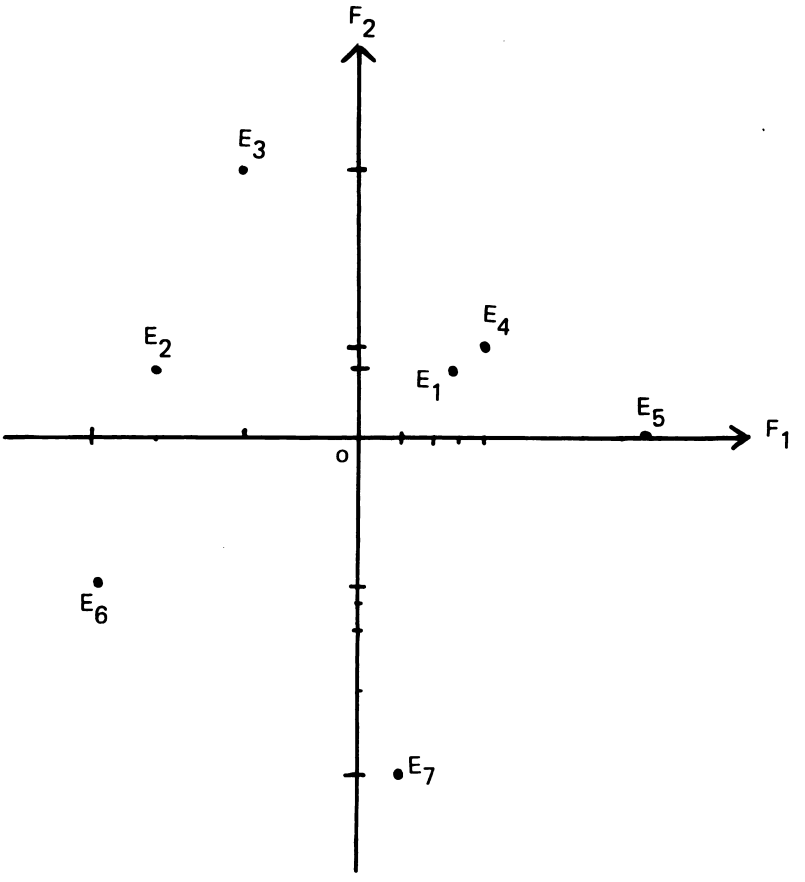
$$F_1\ (4) = .640811$$

Valor que asume el factor 1 para la empresa 4.

A los efectos de una primera aproximación a la conformación de grupos con estas siete empresas, graficamos en un par de ejes (factores) tomando como coordenadas los valores presentados en la matriz anterior.

---

* El término "multiplicado" se utiliza aquí como sinónimo de producto escalar entre dos vectores. Es decir, se efectúa el producto de los números que ocupan igual posición en ambos vectores y se suman los resultados.



Podrían realizarse diferentes tipos de agrupamientos, de acuerdo al objetivo perseguido. Por ejemplo, podrían agruparse de acuerdo al valor que adopta el factor dos para cada empresa. De hacerse esto, las empresas 1, 2, 3 y 4 formarían un grupo y las 6 y 7 formarían otro grupo. Quedaría la empresa 5 sin clasificar.

Si se adopta el criterio de asignar la empresa 5 a aquel grupo cuyo centro diste menos de ella, evidentemente, esta empresa formaría parte del primer grupo.

2.4 Referencias

1 ANDERSON, T.W., "Introduction to Multivariate Statistical Analysis", John Wiley, New York, 1958.

2 CATTELL, R.B., Factor Analysis: An Introduction to Essentials. (I) The Purpose and Underlying Models, (II) The Role of Factor Analysis in Research, Biometrics 21: 190 - 215, 405 - 435, 1965.

3 CORDONNIER, CARLES y MARSAL, "Economía de la Empresa Agraria", Ediciones Mundi-Prensa, 1973.

4 FINKBEINER, D.T., "Introduction to Matrices and Linear Transformations", W.H. Freeman and Company, 1966.

5 GRAYBILL, F.A., "An Introduction to Linear Statistical Models", Mc Graw - Hill, 1961.

6 HAGOOD, M. and BERNERT, E., "Component indexes as a basis for Stratification in Sampling". Journal of the American Statistical Association, 40, 330 - 341, 1945.

7 I.I.C.A., "Seminario sobre Métodos y Problemas en Tipificación de Empresas Agropecuarias", Serie de Informes de Conferencias, Cursos y Reuniones No. 92, Volúmenes 1, 2 y 3, Montevideo, 1975.

8 LEBART, L. et FENELON, J.P., "Statistique et informatique apliqués", Dunod, 1975.

9 LINDGREN, B.W., "Statistical Theory", Macmillan, 1962.

10 MORRISON, D.F., "Multivariate Statistical Methods", Mc Graw - Hill, 1967.

11 TINTNER, G., "Econometrics", John Wiley, New York, 1952.

12 WILKS, S.S., "Mathematical Statistics", John Wiley & Sons, Inc., 1962.

CAPITULO 3

Algunas técnicas de conglomeración. Su naturaleza y sus posibilidades en tipificación de empresas.

Algunas técnicas de conglomeración. Su naturaleza y sus posibilidades en tipificación de empresas.

3

Alfredo Alonso — DIEA

3.1 Resumen general

Se presentan en este trabajo una serie de técnicas de clasificación, que ofrece el Análisis de Conglomeración, con referencia a la tipificación de empresas agropecuarias. En especial se describen los algoritmos de cómputo de Van Rijsbergen, Ward y Sparks, que han sido programados en cooperación con el Instituto Interamericano de Ciencias Agrícolas.

Se describen luego en forma sucinta las técnicas de Análisis Discriminante, Tablas de Contingencia y Dócima de Kruskal y Wallis, haciendo referencia a la forma en que éstas pueden ser utilizadas para probar la calidad de los agrupamientos obtenidos mediante la aplicación de los métodos de Análisis de Conglomeración.

No se hacen indicaciones definitivas sobre cuáles son las técnicas más recomendables para utilizar en problemas de tipificación, debido a que no se dispone de suficientes antecedentes sobre aplicaciones realizadas con estos métodos.

De acuerdo con esto, se entendió conveniente presentar distintos métodos, tanto para clasificar como para analizar las clasificaciones a posteriori. Al comparar los resultados obtenidos mediante la aplicación de metodologías alternativas, se podrán definir con mayor claridad las empresas tipo y se irán acumulando experiencias que permitirán ir ajustando una metodología más precisa.

Desde el Seminario realizado en noviembre de 1975, hemos podido progresar con apoyo del IICA en operacionalización de técnicas novedosas en el tema y en generación de primeras experiencias sobre casos concretos. El camino en realidad, recién se inicia.

3.2 Introducción

Cuando se encara un problema de tipificación de empresas agropecuarias, se supone que existen una serie de características o atributos que permitirán extraer las diferencias y/o semejanzas que se presentan entre las explotaciones, a partir de las cuales se puede proceder a realizar agrupamientos.

Los sistemas más antiguos y difundidos de clasificación se basan en el tamaño de las explotaciones y en el valor bruto de producción, agrupando las empresas de acuerdo con su tamaño o con el rubro o combinación de rubros que genera la mayor parte del ingreso bruto (17).

Todo criterio de clasificación debería considerar el mayor número posible de atributos que resulten distintivos de los elementos que se quieren agrupar. Este concepto se puede aplicar, también, a las empresas agropecuarias que presentan diferencias apreciables, no sólo en características físicas, sino también en ciertos aspectos que resultan más difíciles de cuantificar como ser la capacidad empresarial de los productores.

Una clasificación basada exclusivamente en una o dos variables (del tipo "valor bruto de producción" o "tamaño de las explotaciones") presenta, entonces, el inconveniente de que se hace un uso muy pobre de la información disponible. Se toman en cuenta pocas características y, por más importantes que éstas sean, se dejan de lado otros atributos que pueden ser relevantes y que podrían agregar información útil para lograr una mejor comprensión del conjunto de empresas que se quiere clasificar.

Se presenta, entonces, el Análisis de Conglomeración ("cluster analysis") que permite resolver el problema de la tipificación de empresas agropecuarias en base a un número elevado de variables.

En el capítulo 3.3 de este trabajo, se indican algunas de las técnicas que pueden ser utilizadas para clasificar empresas.

En el capítulo 3.4 se presentan técnicas estadísticas que permitan analizar los conglomerados obtenidos al aplicar los métodos a que se hizo referencia en el capítulo que lo precede.

El presente trabajo se ha encarado como un aporte al uso de las metodologías propuestas por Ferreira, P. y Kaminsky, M., en el "Seminario sobre métodos y problemas en tipificación de empresas agropecuarias"*, que se desarrolló en Montevideo en 1975. De acuerdo con esto, se supone el conocimiento del material que estos autores presentaron en esa oportunidad.

3.3 Métodos de clasificación

En este capítulo se presentan algunos métodos de clasificación que pueden ser utilizados para agrupar empresas, como paso previo a la definición de empresas tipo. Dentro de las múltiples técnicas que ofrece el Análisis de Conglomeración, se hace referencia a las relacionadas con los algoritmos de cómputo que han sido programados para el IICA (12).

Mediante la aplicación del Análisis de Conglomeración se pueden agrupar las explotaciones en conglomerados o "clusters" tales que las diferencias entre los elementos que forman parte de un conglomerado sean mínimas y la diferencia entre conglomerados sea máxima.

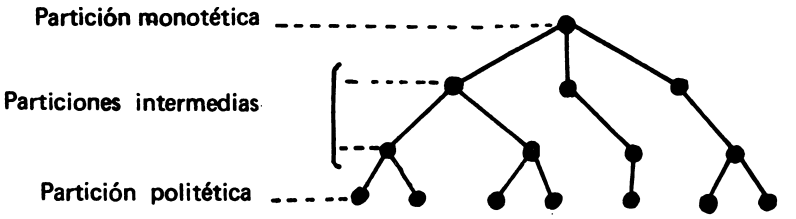
Entre las múltiples técnicas que ofrece la taxonomía matemática para realizar este tipo de análisis y más precisamente dentro de las técnicas estadísticas, Harrison, I. (7), establece una distinción entre las técnicas algorítmicas y las heurísticas: "técnicas algorítmicas son aquéllas que garantizan llegar a una solución que es óptima en un sentido previamente definido y técnicas heurísticas son las que ayudan al descubrimiento e interpretación de hechos y verdades, pero que sólo proporcionan soluciones que son buenas, generalmente en un sentido no tan bien definido".

Las técnicas heurísticas son las que se utilizan con mayor frecuencia por ser de mayor aplicación práctica y porque, en realidad, es muy difícil definir una partición globalmente óptima de un conjunto dado.

---

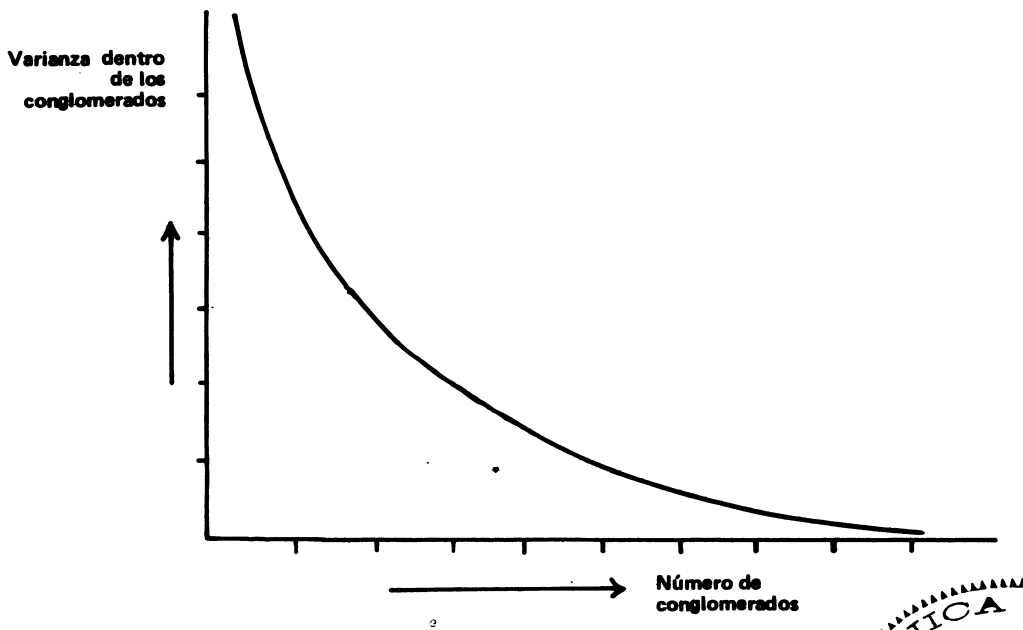
* Véase (8)

Al aplicar técnicas heurísticas, no surge del procedimiento cuáles y cuántos son los conglomerados que nos dan una partición óptima, se obtiene con ellas solamente una jerarquía de agrupamientos que cubre casos desde la partición monotética (todas las empresas en un cluster), hasta la partición politética (cada empresa es un cluster).



Algunos procedimientos de conglomeración son jerárquicos en el sentido de que los elementos se van agrupando, en forma secuencial, de modo tal que dos elementos que se unen en cualquier etapa pasan a constituir una unidad que no se separará hasta finalizar el proceso clasificatorio. Así, en cada paso se obtiene una partición que es, en cierto modo, la mejor para ese nivel. De acuerdo con los objetivos de la clasificación y en base a criterios más o menos arbitrarios, se debe optar por alguna de las particiones intermedias.

A medida que se reduce el número de conglomerados, los elementos agrupados van siendo cada vez menos homogéneos, de modo que el investigador debe balancear la ventaja de trabajar con un número reducido de empresas tipo, frente al hecho de que éstas surjan como promedio de un conjunto cada vez más heterogéneo de explotaciones.



Al ser mayor el número de conglomerados, va disminuyendo la varianza dentro y aumenta la varianza entre clusters.

Siguiendo a Harrison, I. (7), podemos hacer una distinción entre las técnicas de conglomeración, separando las técnicas estadísticas elementales de las más complejas. De acuerdo a esto, se describen en las secciones siguientes algunas técnicas de clasificación con referencia a los algoritmos de cómputo que han sido programados en el IICA y que se encuentran disponibles para ser utilizados.

Previamente, se hace una breve reseña de algunos métodos para calcular distancias, debido a que varios de los métodos de análisis de conglomeración operan a partir de la formulación de medidas de disimilaridad entre las observaciones.

3.3.1 Métodos para calcular distancias

Cuando se utilizan métodos de conglomeración basados en la formulación de una matriz de distancias entre observaciones, se debe tener en cuenta que la validez de la clasificación resultante va a estar determinada por la forma en que estas distancias se calculan.

Al construir las medidas de disimilaridad entre explotaciones, se está resumiendo la información que se posea sobre las empresas, por lo que se debe elegir, como lo indica Ferreira, P. (8), una métrica que minimice esta inevitable pérdida de información.

Para construir una matriz de distancias "D", entre "n" establecimientos, a partir de la información contenida en "m" variables explicativas, se puede formular como una primera aproximación la **distancia euclideana** entre empresas:

$$d(i,j) = \sqrt{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2} \; ; \quad i,j = 1,n \; ; \; i < j$$

Es conveniente normalizar los valores de las variables antes de construir las distancias, para que el peso relativo de cada variable no quede determinado esencialmente por las unidades de medida.

En la determinación de las distancias se pueden proponer atributos que deben pesar más que otros, por ser más relevantes, de acuerdo con los objetivos de la clasificación ensayada. A estos efectos se puede encarar la ponderación de las variables, asignándoles pesos relativos W_k diferenciales. La distancia entre los establecimientos i y j sería:

$$d(i,j) = \sqrt{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2 \cdot W_k} \; ; \quad i,j = 1,n \; ; \; i < j$$

Se plantea así la asignación de ponderaciones a las variables, lo que lleva a incluir más elementos subjetivos, que van a afectar en forma significativa el resultado del análisis. Sin embargo, si no se ponderan explícitamente, lo que se hace es ignorar el problema y no resolverlo, en la medida en que eso implica asignarle igual peso a todas las variables.

Otro aspecto que se debe considerar, es que las variables que se utilizan para construir las medidas de disimilaridad se encuentran generalmente correlacionadas. Esto determina que variables altamente correlacionadas van a pesar más al definir las distancias. Por ejemplo: si se toman variables como superficie total y superficie dedicada a pastoreo, el concepto de "tamaño de la explotación" se estará considerando dos veces en la formulación de las distancias entre empresas.

Una forma de solucionar este problema es la ensayada por Green et al (6), quienes realizaron un análisis de componentes principales sobre las variables originales y luego construyeron las medidas de disimilaridad a partir de los valores de las dos primeras componentes, que explicaban el 74% de la varianza total. El inconveniente que surge de aplicar este procedimiento es la pérdida de información que se produce en la medida que se retiene un poder explicativo de la varianza inferior al 100%.

Para evitar el hacer un análisis de componentes principales, y tomando en cuenta el 100% de la varianza total, se pueden construir medidas de distancias que compensan por la correlación entre las variables. Tal es el caso de la **métrica de Mahalanobis**, que se define:

$$d(i, j) = \sqrt{(X_i - X_j)' \; V^{-1} \; (X_i - X_j)}$$

donde X_i, X_j son los vectores que contienen el valor de las variables para los establecimientos i y j y V es la matriz de varianzas-covarianzas entre las variables. Las variables se pueden ponderar de acuerdo con la importancia relativa que se les asigna, para lo que se debe construir una matriz diagonal W con los pesos relativos. La distancia queda entonces planteada en la siguiente forma:

$$d(i, j) = \sqrt{(X_i - X_j)' \; W' \; V^{-1} \; W \; (X_i - X_j)}$$

Morrison, D.G. (14), hace un breve análisis de las propiedades de varios métodos para calcular distancias y se puede considerar acertada su conclusión de que la métrica de Mahalanobis con ponderaciones es, en general, la más recomendable.

Mediante los programas que se encuentran disponibles en el IICA, se pueden calcular las distancias euclideana y de Mahalanobis y el usuario tiene la opción de estandarizar y ponderar las variables si lo desea.

Se debe tener presente que las medidas de disimilaridad que se describieron en esta sección son sólo algunas de las métricas que pueden ser utilizadas. Por ejemplo, en un trabajo de CEPAL (2), se propone otra que permite también eliminar el problema de variables correlacionadas, mediante la corrección de Ivanovic.

Se creyó conveniente con el IICA, trabajar en una primera etapa con un número no muy elevado de opciones de distancias, para ir introduciendo nuevas medidas, luego de comparar las que se han definido en esta sección.

Después de revisar las métricas que pueden utilizarse, pasamos a describir los métodos de conglomeración.

3.3.2 Técnicas estadísticas elementales

Estos métodos de conglomeración se aplican a partir de la matriz de distancias (D) entre observaciones y se plantean en base a una representación gráfica en la cual cada observación se representa por un punto. Estos puntos van siendo, luego, enlazados de acuerdo con las distancias calculadas.

Algunos de los métodos que se basan en estos diagramas son el de enlace singular (single link), propuesto por Florek y Sneath (5), el de enlace completo (complete link), propuesto por Sorenson y los métodos de enlace promedio de Sokal y Michener.

i) Algoritmo de Van Rijsbergen

Este algoritmo (18), trabaja de acuerdo con el método de enlace singular, por el cual los puntos (explotaciones en el caso que nos interesa fundamentalmente), se van uniendo a medida que aparecen en la matriz D valores menores o iguales a un nivel prefijado (h) para realizar el análisis.

Al terminar el análisis, cada subgráfico formado por puntos enlazados, se define como un agrupamiento, de modo que se obtiene una partición del conjunto original en subconjuntos, algunos de los cuales pueden ser singulares (una sola empresa).

El algoritmo trabaja a partir de un nivel "h" prefijado por el analista, y en cada paso lo va incrementando en un valor dado.

Al ir aumentando el valor de "h", aparecerá en cada iteración un mayor número de enlaces entre observaciones y se irá reduciendo el número de conglomerados, de modo que el método de clasificación es jerárquico y de tipo aglomerativo.

Los agrupamientos que se obtienen por este método, deben ser analizados cuidadosamente, porque se pueden formar cadenas de puntos donde la distancia entre los extremos de la cadena es muy grande. Asimismo, se debe encarar el problema de asignar las observaciones que quedan aisladas al finalizar el análisis. Un criterio que se puede adoptar es el de calcular las distancias de cada elemento aislado con respecto a los centros de los conglomerados formados y asignarlos de acuerdo con la menor, o sea al cluster más cercano.

ii) Eficiencia de las clasificaciones obtenidas

Para medir el grado de eficiencia relativa de las clasificaciones obtenidas, Van Rijsbergen propone un índice que relaciona las distancias entre los elementos que forman parte de los conglomerados, con el coeficiente de disimilaridad ultramétrico que caracteriza la partición. Lo podríamos llamar "índice de eficiencia relativa" (IER) y tiene la siguiente forma:

$$IER = \frac{\Sigma\, d\,(A,B) \;-\; \Sigma\, D\,(d)\,(A,B)}{\Sigma\, d\,(A,B)}$$

donde d(A,B) son las distancias entre las observaciones que forman parte de cada cluster y D(d) (A,B) es igual al mínimo nivel "h" para el cual las observaciones A, B pertenecen al mismo conglomerado.

El algoritmo disponible estima en cada iteración el valor de Σ D(d) (A,B) para el valor "h" que caracteriza la partición en la siguiente forma:

- va calculando para cada conglomerado el número de distancias posibles entre los elementos que lo forman
- multiplica por el valor "h" y acumula.

De modo que:

$$\Sigma\ D(d)\ (A,B) = \sum_{Sk=1}^{g} C_2^{nk} \times h$$

donde nk: número de elementos en el cluster Sk (Sk = 1, g)

El IER se puede entonces calcular fácilmente, ya que Σd(A,B) se obtiene directamente de la matriz de distancias original, mediante la suma de las distancias entre los elementos que forman cada conglomerado, o sea:

$$\Sigma\ d\ (A,B) \quad = \sum_{Sk=1}^{g} \quad \sum_{A<B\ \epsilon\ Sk} \quad d\ (A,B)$$

El IER resulta similar al índice de clasificabilidad (IC) reportado por Kaminsky, M. (8);

$$IC \quad = \quad \frac{\sum_{i<j} d^*(i,j)}{\sum_{i<j} d\ (i,j)}$$

definiendo d*(i, j) en la misma forma que D(d) (A,B) y d(i, j) igual que d(A,B). Sin embargo, al hacer la suma con i < j, el numerador variará únicamente con el valor de "h", sin tener en cuenta el número de elementos agrupados.

Podemos considerar, entonces, que el IER planteado por Van Rijsbergen, dará una idea más ajustada de la homogeneidad de los clusters formados, en la medida en que compara las distancias reales entre los elementos que forman un conglomerado con la distancia "h" que se tomó para construirlos.

Si se obtienen conglomerados en forma de cadena, con valores d(A,B) entre los elementos muy superiores al valor "h", el valor del índice será alto, mientras que se obtendrán valores negativos cuando las observaciones que forman un conglomerado disten entre sí, en promedio, menos que el valor "h".

3.3.3 Técnicas estadísticas más complejas

Para proceder a la clasificación de los elementos de un conjunto, estos métodos se basan en la minimización de la varianza dentro de los conglomerados.

Si tenemos una partición en dos subconjuntos de un conglomerado dado, la varianza total del conjunto es igual a la suma de las **varianzas dentro** de los subconjuntos disjuntos que lo forman más la **varianza entre** los subconjuntos.

Teniendo en cuenta que la varianza es igual a la suma de cuadrados (desviaciones con respecto a la media al cuadrado) dividida por el número de observaciones, nos manejaremos en adelante con la suma de cuadrados (SC) como estimador de la varianza.

Si se parte de un conjunto A formado por dos subconjuntos disjuntos, A_1 y A_2, compuestos por n_1 y n_2 elementos y $\bar{X}_1$, $\bar{X}_2$ y $\bar{X}$ son las medias aritméticas de los conjuntos A_1, A_2 y A respectivamente, se verifica que:

$$\sum_{X\ \epsilon\ A} (X-\bar{X})^2 = \sum_{X\ \epsilon\ A_1} (X-\bar{X}_1)^2 + \sum_{X \epsilon A_2} (X-\bar{X}_2)^2 + n_1\ (\bar{X}_1 - \bar{X})^2 + n_2\ (\bar{X}_2 - \bar{X})^2$$

o sea que la suma de cuadrados total es igual a la suma de cuadrados dentro de los subconjuntos A_1 y A_2, más la suma de cuadrados entre los subconjuntos A_1 y A_2.

Para dividir un conjunto en dos subconjuntos que sean lo más homogéneos posibles, se debe proceder a minimizar la S.C. dentro de los agrupamientos, lo que llevará maximizar la S.C. entre conglomerados.

Como lo indican Edwards y Cavalli-Sforza (4), la S.C. dentro de un conjunto (intravarianza) es igual a la suma de las distancias al cuadrado entre todos los pares de elementos que lo componen (tomando cada distancia una sola vez), dividida por el número de elementos que componen el conjunto.

Se puede proceder, entonces, a agrupar elementos con el criterio de minimizar la intravarianza a partir de la matriz de distancias enclideanas al cuadrado (D^2).

Pará facilitar la demostración se puede tomar un conjunto formado por dos elementos y una sola variable, considerando que, como se trabaja con distancias al cuadrado, esto es válido para cualquier número de variables y elementos.

Sea un conjunto A formado por dos elementos, A_1 y A_2.

Entonces:

$$\text{S.C. (dentro A)} = (XA_1 - \bar{X})^2 + (XA_2 - \bar{X})^2$$

$$= XA_1^2 + \bar{X}^2 - 2XA_1\bar{X} + XA_2^2 + \bar{X}^2 - 2XA_2\bar{X}$$

$$= XA_1^2 + XA_2^2 + 2\bar{X}^2 - 2\bar{X}(XA_1 + XA_2)$$

$$= XA_1^2 + XA_2^2 + 2\bar{X}^2 - 2\bar{X}(2\bar{X})$$

$$= XA_1^2 + XA_2^2 - 2\bar{X}^2$$

$$= XA_1^2 + XA_2^2 - \frac{(XA_1 + XA_2)^2}{2}$$

$$= XA_1^2 + XA_2^2 - \frac{XA_1^2 + XA_2^2 + 2XA_1XA_2}{2}$$

$$= \frac{2XA_1^2 + 2XA_2^2 - XA_1^2 - XA_2^2 - 2XA_1XA_2}{2}$$

$$= \frac{XA_1^2 + XA_2^2 - 2XA_1XA_2}{2}$$

$$= \frac{(XA_1 - XA_2)^2}{2}$$

$$= \frac{d^2(A_1, A_2)}{2}$$

$$\boxed{\text{S.C. (dentro A)} = \frac{\Sigma\, d^2(i, j)}{n}}$$

De los métodos que se basan en la minimización de la varianza dentro de los conglomerados, han sido programados con el IICA (12), el algoritmo de Sparks (15) y el algoritmo de Ward, con las alternativas propuestas por Lance y Williams, según lo expuesto por Wishart, D. (19).

i) Algoritmo de Ward

El algoritmo de Ward es jerárquico y de tipo aglomerativo; a partir de la partición politética va agrupando elementos o conglomerados para llegar a la partición monotética.

En cada iteración se consideran todas las uniones posibles entre conglomerados y se elige la que produce el menor incremento en la suma de cuadrados dentro de clusters.

Se trata, entonces, de agrupar minimizando la función objetivo F definida como sigue:

$$F = \sum_{t=1}^{T} F_t$$

donde F_t es la S.C. dentro del conglomerado t (t = 1,T)

El incremento en la función objetivo que se produce al unir los conglomerados S_p y S_q para formar un nuevo cluster S_r será:

$$I_{pq} = F_r - F_p - F_q$$

y Ward demuestra (19) que este incremento es:

$$I_{pq} = \frac{k_p \; k_q}{k_r} \; d^2(p,q)$$

donde k_p, k_q y k_r son el número de elementos en los clusters S_p, S_q y S_r respectivamente.

En cada iteración se fusionan los conglomerados S_p y S_q que producen el menor incremento en la función objetivo.

El algoritmo trabaja a partir de la matriz de distancias al cuadrado y en un primer paso une a los dos elementos más cercanos, ya que busca la menor distancia d^2 (i, j).

La S.C. dentro de clusters o función objetivo F es nula al comenzar el análisis, debido a que se parte de la partición politética, y el incremento que se produce al formar el primer conglomerado S_r es igual a la mitad de la distancia al cuadrado entre las observaciones, o sea:

$$I_{pq} = \frac{k_p \; k_q}{k_r} \; d^2 (p, q)$$

$$= \frac{1 \times 1}{2} \; d^2 \; (p, q)$$

$$\boxed{I_{pq} = \frac{1}{2} \; d^2 \; (p, q)}$$

A partir de esta unión, el algoritmo procede a corregir las distancias de los demás elementos con respecto al cluster S_r recién formado, de forma tal que las nuevas distancias d^2i,r ($i = 1, n;\ i \neq p \neq q$) no se deben interpretar más en el sentido usual, sino que van a expresar el doble del incremento que se producirá al unir dos conglomerados, o sea:

$$d^2 (i, j) = 2\, I_{i,j}$$

Se continúa el proceso de la misma forma, eligiendo la menor d^2 (i, j) para ver qué conglomerados se deben unir e incrementando la función F en 1/2 de $d^2(i, j)$, de modo que al realizar n – 1 iteraciones, se llega a la partición monotética.

Al finalizar el análisis, la función objetivo dará el valor de la S.C. total. Esto es así porque, al tener un solo conglomerado, la varianza dentro de él es igual a la varianza total y la varianza entre conglomerados es nula.

La transformación de las distancias luego de cada fusión se realiza mediante la fórmula:

$$d^2 (i,r) = a_p\, d^2 (i,p) + a_q\, d^2 (i,q) + \beta\, d^2 (p,q) + \partial\, [\, d^2 (i,p) - d^2 (i,q)\,]$$

con: $$a_p = \frac{k_i + k_p}{k_i + k_r}$$

$$a = \frac{k_i + k_q}{k_i + k_r}$$

$$\beta = \frac{-k_i}{k_i + k_r}$$

$$\partial = 0$$

y d^2 (i,r) es la distancia al cuadrado entre el conglomerado S_i y el conglomerado S_r ($S_r = S_p \cup S_q$).

El programa de Wishart adaptado para el IICA, permite emplear métodos alternativos al expuesto para corregir las distancias. Lance y Williams (9), los definen como combinatorios, en el sentido de que las nuevas distancias que se van calculando surgen como una combinación lineal de las distancias originales.

Estos métodos, contrariamente a lo que ocurre con Ward, con compatibles debido a que las nuevas distancias que se calculan en cada iteración son de la misma naturaleza que las contenidas en la matriz original D^2.

Mediante la fórmula general expuesta anteriormente y cambiando el valor de los parámetros a_p, a_q, β y ∂ se definen estos métodos en la siguiente forma:

- **Centroide** $\quad a_p = \frac{k_p}{k_r}\ ;\quad a_q = \frac{k_q}{k_r}\ ;\quad \beta = a_p\, a_q\ ;\quad \partial = 0$

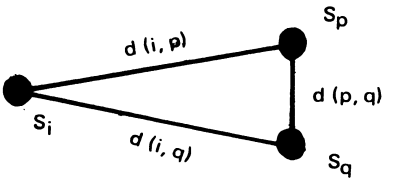
- **Mediana** $a_p = a_q = \frac{1}{2}$; $\beta = -\frac{1}{4}$; $\partial = 0$

- **Promedio de grupo** $a_p = \frac{k_p}{k_r}$; $a_q = \frac{k_q}{k_r}$; $\beta = \partial = 0$

- **Vecino más cercano** $a_p = a_q = \frac{1}{2}$; $\beta = 0$; $\partial = -\frac{1}{2}$

- **Vecino más lejano** $a_p = a_q = \frac{1}{2}$; $\beta = 0$; $\partial = \frac{1}{2}$

A continuación se presenta una explicación somera de qué significan estas propuestas alternativas. Para facilitar la comprensión del funcionamiento de los tres métodos anotados en primer término, se hace una representación gráfica, pero se debe recordar que en realidad los métodos trabajan con distancias al cuadrado:



Centroide: Este método trabaja sobre la noción usual de distancias, calculando las medidas de disimilaridad entre los centros de los clusters que se van formando.

En la primera iteración calcula las distancias al cuadrado de cada observación S_i (conglomerados singulares) con el centro del cluster reción formado S_r ($S_r = S_p$ u S_q). Así continúa sucesivamente, tomando los centros de los clusters para calcular las d^2 (i, r).

El incremento en la función objetiva F se debe calcular expresamente en cada iteración mediante la fórmula expuesta anteriormente:

$$I_{pq} = \frac{k_p \; k_q}{k_r} \; d^2(p,q)$$

Operando en esta forma, se obtendrá al finalizar el análisis el mismo resultado que con Ward, o sea, la S.C. total.

Una desventaja que se le ha señalado a este método es que las características de los grupos formados por pocos elementos se pierden, debido a que el centroide se ubica siempre más cerca del grupo que tiene el mayor número de elementos.

Mediana: Trabaja en forma similar al centroide, pero sin considerar el tamaño de los grupos. El centro del conglomerado que surge de una fusión estará ubicado en el punto medio del lado más corto del triángulo que se representó más arriba, y la distancia d(i, r) se sitúa sobre la mediana del triángulo. De este hecho es que deriva el nombre del método.

Promedio de grupo. Calcula las distancias en forma similar al centroide pero sin considerar la distancia entre los clusters S_p y S_q simplemente pondera las distancias al cuadrado a los clusters S_p y S_q de acuerdo con el número de elementos que los componen.

Vecino más cercano. Toma la distancia de cada observación S_i con el clusters S_r formado, como la menor de las distancias al cuadrado entre el elemento S_i con los elementos S_p y S_q, o sea:

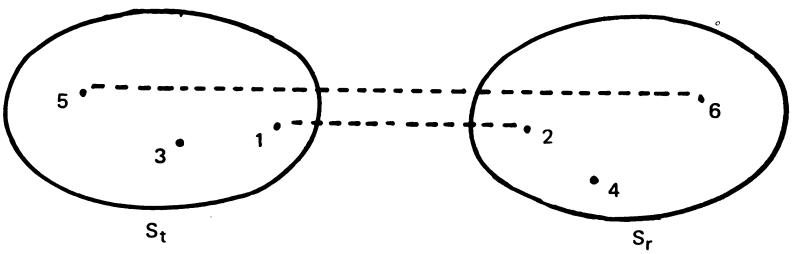
$$d^2(i,r) = \min \left\{ d^2(i,p);\ d^2(i,q) \right\}$$

La medida de disimilaridad entre dos conglomerados, será entonces la distancia al cuadrado entre las dos observaciones $i \in S_i$ y $r \in S_r$ más cercanas.

Vecino más lejano. Opera en forma inversa al precedente, de modo que la distancia de cada observación S_i con el clusters S_r formado será:

$$d^2(i,r) = \max \left\{ d^2(i,p);\ d^2(i,q) \right\}$$

Por ejemplo, si tenemos dos conglomerados S_t y S_r, las distancias calculadas mediante los dos métodos serán:



Vecino más cercano: $d^2(t,r) = d^2(1,2)$

Vecino más lejano: $d^2(t,r) = d^2(5,6)$

Siguiendo a Lance y Williams (9) podemos dividir a estos métodos en conservadores y distorsionadores del espacio original.

Las medidas de disimilaridad contenidas en la matriz original definen un espacio que tiene determinadas propiedades, pero cuando los grupos comienzan a formarse mediante la aplicación de estos métodos de conglomeración, las nuevas distancias que se van calculando definen un nuevo espacio que puede tener o no las propiedades originales.

Al utilizar métodos como Centroide, Mediana y (en forma menos rigurosa) Promedio de Grupo, se tiene que las propiedades del espacio original se mantienen incambiadas, de modo tal que se denomina a estos métodos como conservadores del espacio.

Al aplicar otros métodos sucede que el espacio en la cercanía de los grupos formados se ha dilatado o contraído, de modo que se los puede calificar como distorsionadores del espacio.

El método de Vecino más Cercano contrae el espacio original, de modo que los elementos tienden a unirse a grupos ya formados en lugar de constituirse en centros de nuevos clusters.

El método de Vecino más Lejano y en menor medida el método de Ward, actúan como dilatadores del espacio original y al aplicarlos los elementos tienden a actuar preferentemente como núcleos de nuevos grupos.

Los autores citados aconsejan, en forma general, la utilización combinada de un método conservador del espacio y un método levemente dilatador.

En el IICA se encuentra programado el algoritmo general, con opción para el usuario de elegir cualquiera de las seis transformaciones. Operándolo se obtiene en cada paso la cantidad de elementos que constituyen cada conglomerado y el número de cluster al cual pertenece cada observación. Para el método de Ward se obtiene además el valor de la función objetivo F.

Se tiene poca experiencia sobre la aplicación de los distintos criterios, pero aparecen como más aconsejables los métodos conservadores del espacio como Centroide y Promedio de grupo o levemente dilatadores como Ward, que permite además obtener en forma directa una estimación de la varianza dentro de grupos.

ii) Algoritmo de Sparks

A partir de una matriz de observaciones en n variables, los elementos son agrupados en conglomerados de modo que la suma de cuadrados dentro de los agrupamientos sea mínima.

Al iniciar el análisis, el usuario debe fijar el número de clusters y los centros iniciales de los conglomerados. Para cada observación, el programa calcula las distancias a los centros iniciales y las asigna al más cercano. Luego calcula la media de los agrupamientos formados y estos valores pasan a ser los nuevos centros.

Se examinan, entonces, todas las observaciones para ver si al cambiarlas hacia otro conglomerado se produce una disminución en la suma de cuadrados dentro de clusters.

Se reasigna una observación que pertenece al cluster S_k sólo si la suma de los desvíos al cuadrado con respecto al centro del cluster S_t es menor que con respecto al centro de S_k, aún cuando los centros cambien simultáneamente, o sea, cuando:

$$\frac{n_t}{n_t + 1} \; d_t^2 < \frac{n_k}{n_k - 1} \; d_k^2$$

donde n_t: número de observaciones en S_t

d_t^2: distancia de la observación al centro de S_t.

El procedimiento se sigue repitiendo hasta que no se produce ninguna reasignación. Al finalizar el análisis, se obtienen los centros finales de los clusters, el número de cluster al cual pertenece cada observación, la suma de cuadrados dentro de cada conglomerado y el número de observaciones en cada agrupamiento.

Sparks señala que los centros iniciales deben ser elegidos cuidadosamente para reducir el tiempo de computación y para evitar el obtener óptimos locales en lugar del óptimo global. A partir de la experiencia acumulada con la aplicación de este algoritmo podemos señalar que trabajando con poblaciones que no presentan grupos perfectamente definidos el problema de la obtención de mínimos locales se torna sumamente significativo. En efecto, al cambiar los centros iniciales se obtienen agrupamientos diferentes, de modo que se hace difícil precisar cuando se ha llegado al óptimo global.

Además, como este método no es jerárquico, aparece como recomendable realizar la comparación de varias particiones, para cada una de las cuales se debe probar con diferentes centros iniciales. De esta forma se podrá tener una idea más acertada de la robustez de los agrupamientos resultantes.

Para elegir los centros iniciales se puede hacer una distribución de frecuencias de las observaciones, de acuerdo con el valor que presentan para alguna variable que se considera relevante. Haciendo tantos tramos como grupos se piensan obtener, se puede elegir una empresa de cada clase para proponerla como centro inicial, con cierta confianza de que se podrá reducir el tiempo de computación.

El "Euclidean Cluster Analysis" de Sparks, trabaja con distancias euclideanas al cuadrado, entre las observaciones y los centros de los clusters, por lo que previamente a su aplicación se debe decidir sobre la estandarización y ponderación de las variables.

El método no permite corregir por la correlación entre los atributos, de modo que si se trabaja con variables altamente correlacionadas, podría resultar conveniente aplicarlo a partir del valor de las Componentes Principales que surjan de un análisis previo.

3.3.4 Conclusiones Generales

En las secciones precedentes se han expuesto algunos de los métodos de clasificación que se pueden utilizar para tipificar empresas agropecuarias.

Debemos concluir que no se puede hablar de una técnica de clasificación que permita someter la masa de datos de que se dispone a un proceso computacional del cual van a surgir grupos perfectamente definidos.

El analista debe optar en cada paso, entre una serie de alternativas de acuerdo con su conocimiento sobre la información de que dispone y los objetivos de la clasificación.

Las decisiones que se deben tomar son muchas. Se deben elegir las variables, el método de clasificación, las medidas de distancia si corresponde, la estandarización o no de las variables y su ponderación, para obtener una jerarquía de agrupamientos frente a la cual se debe optar por alguna de las particiones resultantes.

Hacer recomendaciones acerca de cuáles de estas técnicas, incluyendo sus opciones, son las más aconsejables para aplicar en un problema de tipificación de explotaciones agropecuarias es muy arriesgado. Aún no se dispone de ensayos que comparen clasificaciones obtenidas mediante la aplicación de diversos métodos y en el tema de tipificación, en particular, no se tiene mucha información sobre aplicaciones realizadas.

A medida que se vaya trabajando en el tema con estas técnicas, si irá ganando en experiencia y se contribuirá a "acortar la distancia que existe entre la teoría y la práctica", como se señalara en las conclusiones del seminario de 1975 (8).

De acuerdo con esto, será conveniente encarar el problema de tipificación mediante la aplicación de más de uno de los métodos disponibles, de modo que si se obtienen resultados similares a partir de la utilización de metodologías alternativas, se podrá tener un mayor grado de confianza en la robustez de las clasificaciones resultantes. Se irán acumulando, además, experiencias que ayudarán a determinar cuáles

son los métodos que permiten obtener clasificaciones más ajustadas, de acuerdo con la naturaleza de los problemas planteados.

3.4 Análisis a posteriori

En este capítulo se describen, brevemente, algunas técnicas estadísticas que permiten analizar las clasificaciones obtenidas mediante la aplicación de los métodos de Análisis de Conglomeración presentados en el capítulo anterior. A este fin se propone la utilización de los métodos de Análisis Discriminante, Tablas de Contingencia y la Dócima de Kruskal y Wallis.

Como señalan Ling y Killough (11) los métodos de conglomeración producen clusters, existan éstos realmente en la población original o no. En la medida en que los procedimientos heurísticos más usuales no proporcionan una prueba de la bondad de las clasificaciones obtenidas mediante su aplicación, se hace necesario el analizar los resultados a posteriori, aplicando técnicas que permitan docimar la calidad de los agrupamientos resultantes.

3.4.1 Análisis discriminante

El Análisis Discriminante es una técnica que permite describir y clasificar elementos representados por un número elevado de variables.

Como indican Lebart y Fenelon (10), "el problema que se propone resolver el A.D. es el de buscar entre todas las combinaciones lineales de las variables, las que tengan una varianza entre ellas máxima (a fin de exaltar las diferencias entre clases) y una intravarianza mínima (de modo que las clases estén bien delimitadas). Estas combinaciones lineales son las funciones discriminantes".

Existe evidentemente una semejanza entre el análisis discriminante y el análisis de conglomeración, pero se debe tener presente que el primero se debe aplicar sobre clases previamente definidas para determinar las combinaciones lineales que discriminan mejor entre grupos, mientras que el análisis de conglomeración se debe utilizar para construir dichas clases.

Para docimar la hipótesis de que las funciones discriminantes pueden haber surgido al azar, se puede computar el estadígrafo D^{2*} o "distancia generalizada de Mahalanobis" que puede ser usado, asumiendo normalidad en la distribución de las variables, como ji cuadrado con m (g - 1) grados de libertad, donde m es el número de variables y g el número de clases.

Al aplicar sobre los conglomerados formados el programa de análisis discriminante que se encuentra disponible, se obtiene:

i) La media de cada grupo.

ii) La matriz de varianzas-covarianzas o "pooled dispersion matrix"

iii) El valor del estadígrafo generalizado de Mahalanobis.

iv) Las funciones discriminantes; para cada grupo se obtienen los coeficientes de la función discriminante que lo caracteriza.

v) Una evaluación de las funciones de clasificación. Para cada observación se tiene la probabilidad asociada con la función discriminante que le asigna mayor valor y la función que corresponde o sea, la probabilidad de que la observación pertenezca al grupo caracterizado por esa función.

* En el capítulo anterior con D^2 representábamos una matriz de distancias al cuadrado; se debe tener presente el cambio de notación. El estadígrafo D^2 representa ahora un escalar.

El valor máximo que puede alcanzar esta probabilidad es 1, cuando existe certeza absoluta de que la observación pertenece a un grupo dado, y como mínimo 1/g, cuando puede pertenecer indistintamente a cualquiera de los grupos.

Al aplicar el Análisis Discriminante sobre los clusters formados, se tendrá, en caso de haber logrado una buena clasificación, un valor del estadígrafo de Mahalanobis altamente significativo, y las observaciones que pertenecen a un mismo conglomerado presentarán una alta probabilidad asociada con la función discriminante correspondiente a ese agrupamiento.

El análisis discriminante puede ser utilizado también para clasificar empresas y puede contribuir a la interpretación de los conglomerados obtenidos:

a) Aplicando Análisis de Conglomeración sobre una muestra extraída al azar de la población que se quiere clasificar, se puede correr un Análisis Discriminante sobre los clusters obtenidos y luego clasificar el resto de las explotaciones, asignándolas al conglomerado que correspondan, de acuerdo al mayor valor que se obtenga al aplicar las funciones discriminantes, como fue propuesto por Kaminsky, M. (8).

b) Si se construye una tabla con los coeficientes de las funciones discriminantes y las variables utilizadas, se puede interpretar, como señala Press, J. (16), la naturaleza de los grupos formados y se verá qué tipo de empresas tenderán a ser clasificadas en cada uno de ellos.

3.4.2 Tablas de contingencia

Las tablas de contingencia permiten docimar la existencia de asociación o dependencia entre dos características o atributos relativos a las unidades de una población dada.

Si suponemos n individuos clasificados según los criterios A y B, de modo que se tienen r clasificaciones en A y s clasificaciones en B y que el número de individuos que pertenecen a A_i y B_j es n_{ij}, se tiene una tabla de contingencia de r x s con frecuencias absolutas n_{ij} en las casillas, de modo que la suma de los n_{ij} sea n.

	B_1	B_2	....	B_s	$n_{i.}$
A_1	n_{11}	n_{12}	....	n_{1s}	$n_{1.}$
⋮	⋮	⋮		⋮	⋮
A_r	n_{r1}	n_{r2}	....	n_{rs}	$n_{r.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	....	$n_{.s}$	n

En la última fila y columna se tienen las frecuencias marginales $n_{i.}$ y $n.j$

La noción de independencia estadística de las dos características, significa el cumplimiento de la relación:

$$n_{ij} = n_{i.} \times n.j$$

o sea que A y B son estadísticamente independientes si y sólo si la frecuencia conjunta para cada celdilla de la tabla es igual al producto de las frecuencias marginales.

Para realizar dócimas de independencia se comparan, en cada celdilla el valor observado (F_o) con el valor esperado teórico en caso de independencia (F_t), que resulta del producto de las frecuencias marginales. Mientras mayores sean las diferencias entre las frecuencias empíricas y teóricas, mayor será el grado de dependencia entre las dos variables. Se construye así un estadístico de contingencia que tiene una distribución próxima a la ji cuadrado con (r-1) (s-1) grados de libertad en la forma:

$$T = \Sigma \frac{(F_o - F_t)^2}{F_t}$$

La dócima se realiza comparando el valor del estadístico calculado con el valor correspondiente tabulado al nivel de significación preestablecido. En la medida en que el valor calculado sea mayor que el valor de tablas, se rechaza la hipótesis nula de ausencia de asociación o independencia.

A partir del valor T calculado, se puede tratar de madir el grado de asociación entre las dos características, mediante algún estedígrafo de contingencia, como el "C" de Pearson, que tiene la siguiente forma:

$$C = \sqrt{\frac{T}{T + N}}$$

donde N es el número de observaciones.

El coeficiente de contingencia es siempre positivo y menor que 1, el límite inferior es cero y el límite superior es, como postula Conover, W.J. (3),

$$\sqrt{\frac{q - 1}{q}} \qquad \text{con } q = \min \{ r , s \}$$

Se tiene, entonces, que las tablas de contingencia permiten probar la existencia de asocieción y el estadígrafo C da idea sobre el gredo de asociación entre las dos características consideradas.

Para analizar clasificaciones obtenidas mediante los métodos de análisis de conglomeración, se pueden plantear tablas de contingencia para docimar la existencia de asociación entre el hecho de que las empresas pertenezcan a un cluster determinado y el valor que presentan para alguna variable.

Debido a que el análisis se debe plantear considerando una sola variable por vez, se podrían hacer tantas tablas como variables se utilizaron en la clasificación o simplemente hacerlo con una variable, que fue utilizada o no en la clasificación, pero que se considera relevante de acuerdo con la naturaleza de las empresas que se quieren tipificar y los objetivos de la clasificación.

Luego se puede calcular algún estadígrafo de contingencia, como el C de Pearson, para tener idea del grado de asociación existente.

3.4.3 Dócima de Kruskal y Wallis

La técnica de Kruskal y Wallis permite docimar la asociación entre dos características o atributos, una de las cuales se divide en categorías, como se hacía en las tablas de contingencia, mientras la otra se presenta mediante un ordenamiento, en la siguiente forma:

	CLASES		
	A	B	C
ORDENAMIENTO	1º	2º	7º
	3º	5º	8º
	4º	6º	11º
	9º	14º	12º
	10º	15º	13º

Se computa un estadístico H que se distribuye como ji cuadrado con (h-1) grados de libertad en la forma:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{h} \frac{R_j^2}{n_j} - 3(N+1)$$

donde: N es el número de observaciones
h es el número de categorías
n_j es el número de observaciones en la categoría j, y
R_j es la suma de los valores de los rangos en la categoría j.

La dócima de asociación entre las categorías consideradas se hace por la comparación del valor de H calculado con el valor de ji cuadrado de tablas al nivel de significación prefijado. En la medida en que el valor de H sea mayor que el valor de tablas se rechaza la hipótesis nula de ausencia de asociación entre las características consideradas.

La dócima de Kruskal y Wallis puede aplicarse de un modo similar a las tablas de contingencia. Se podría docimar la asociación entre el hecho de que una observación pertenezca a un cluster dado frente al ordenamiento de las empresas, de acuerdo con el valor de alguna variable elegida de acuerdo con los objetivos de la investigación.

3.4.4 Conclusiones generales

Se describieron, superficialmente, tres técnicas y la forma en que pueden ser utilizadas para analizar las clasificaciones obtenidas a partir de la aplicación de los métodos expuestos anteriormente.

Evidentemente la técnica que resulta estadísticamente más aconsejable por su robustez es el análisis discriminante, que permite trabajar con todas las variables consideradas para clasificar y obtener una idea ajustada de la bondad, no sólo de los agrupamientos analizados en su conjunto, sino también de la asignación de cada una de las observaciones.

A partir de su aplicación se dispondrá, además, de información sumamente útil para describir las clases consideradas y se podrán clasificar, si existieran, las empresas que perteneciendo a la misma población no habían sido asignadas a los agrupamientos resultantes al aplicar los métodos de análisis de conglomeración.

Las tablas de contingencia y la dócima de Kruskal y Wallis, a pesar de no tener la exactitud del análisis discriminante, pueden contribuir en ciertos casos a la resolución de los problemas de tipificación que se planteen. Por ejemplo, en aquellos casos en que se quiera testear la existencia de relación entre los conglomerados formados y alguna variable, que se usó para clasificar o no, considerada en forma individual, estos métodos pueden resultar de gran utilidad.

De todos modos, siempre resulta conveniente el disponer de metodologías alternativas que puedan ser aplicadas para resolver cualquier problema que se esté investigando. En la medida en que se utilizan en en forma comparativa y se explicitan los resultados, aún en el caso de que sean contradictorios, se le estará dando una mayor validez a las conclusiones que se obtengan.

3.5 Anexo

Se presenta en este anexo un ejemplo numérico con el fin de ilustrar la aplicación de algunas de las técnicas presentadas anteriormente.

Se debe tener en cuenta que dado el bajo número de observaciones no se pueden extraer conclusiones válidas con la mayoría de los métodos utilizados; se pretende únicamente ilustrar sobre su aplicación.

Supongamos que se tienen 6 empresas agropecuarias que se quieren clasificar de acuerdo con la información que aportan 4 variables. Los atributos elegidos de acuerdo con los objetivos de la clasificación son:

1) **S.T.:** Superficie total de la explotación, en hectáreas.
2) **M.O.:** Mano de obra, número de trabajadores por hectárea.
3) **S.C. / S.T.:** Superficie de chacra con respecto a la superficie total.
4) **N.T.:** Un índice que refleja el nivel tecnológico de cada empresa, medido en unidades arbitrarias.

La matriz de observaciones, con los valores que toman las variables para cada empresa y las ponderaciones que se le asigna a cada variable, es la siguiente:

X	S.T.	M.O.	S.C./S.T.	N.T.
1	10	0.20	0.40	300
2	15	0.15	0.36	400
3	45	0.12	0.18	445
4	55	0.10	0.16	900
5	100	0.05	0.14	850
6	150	0.02	0.02	450
W	1.0	1.0	10.0	1.0

A partir de estos datos se pueden ver las clasificaciones que se obtienen al aplicar los métodos de conglomeración:

3.5.1 Algoritmo de Van Rijsbergen

1. Distancia euclideana, sin estandarizar y sin ponderar

i) Matriz de distancias

	1	2	3	4	5	6
1	0	100.12	149.16	601.69	557.31	205.18
2		0	54.08	501.60	457.96	143.96
3			0	455.11	408.72	105.12
4				0	67.27	459.92
5					0	403.11
6						0

ii) Clasificaciones obtenidas

2 Clusters: $\{1, 2, 3, 6\}$ $\{4, 5\}$

3 Clusters: $\{1, 2, 3\}$ $\{4, 5\}$ $\{6\}$

2. Distancia euclideana, estandarizando las variables pero sin ponderarlas

i) Matriz de distancias

	1	2	3	4	5	6
1	0	0.99	2.35	3.71	4.39	5.10
2		0	1.60	2.90	3.53	4.37
3			0	2.02	2.41	2.98
4				0	1.27	3.24
5					0	2.27
6						0

ii) Clasificaciones obtenidas

2 Clusters: $\{1, 2, 3, 4, 5\}$ $\{6\}$

3 Clusters: $\{1, 2, 3\}$ $\{4, 5\}$ $\{6\}$

3. Distancia euclideana, estandarizando y ponderando

i) Matriz de distancias

	1	2	3	4	5	6
1	0	1.25	5.56	6.63	7.40	10.09
2		0	4.42	5.42	6.15	8.93
3			0	2.07	2.58	4.72
4				0	1.35	4.56
5					0	3.56
6						0

ii) Clasificaciones obtenidas

2 Clusters: $\{1,2\}$ $\{3,4,5,6\}$

3 Clusters: $\{1,2\}$ $\{3,4,5\}$ $\{6\}$

4. Distancia de Mahalanobis, estandarizando

i) Matríz de distancias

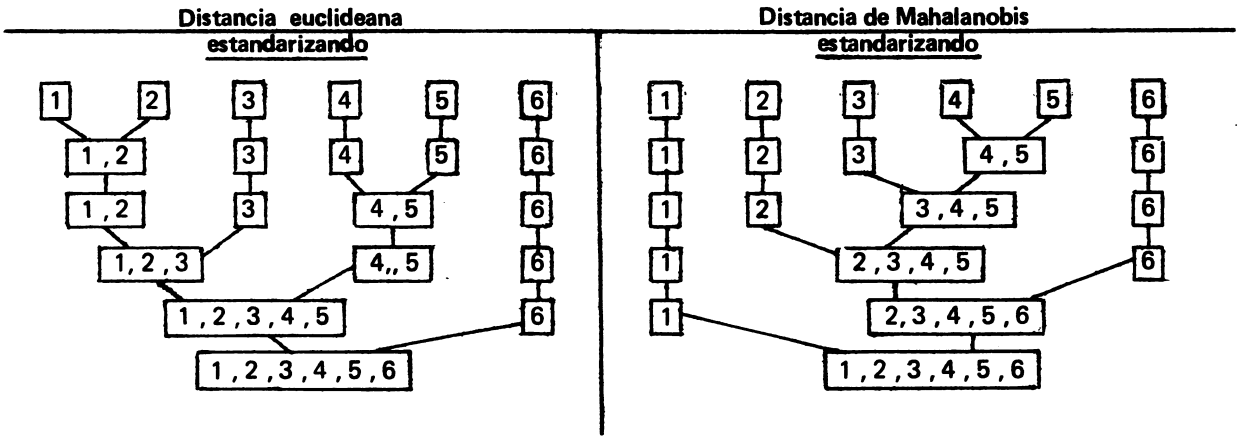
	1	2	3	4	5	6
1	0	1.34	1.37	1.28	1.34	1.35
2		0	1.16	1.41	1.09	1.41
3			0	1.04	1.41	1.18
4				0	0.94	1.40
5					0	1.12
6						0

ii) Clasificaciones obtenidas

2 Clusters: $\{1\}$ $\{2, 3, 4, 5, 6\}$

3 Clusters: $\{1\}$ $\{2, 3, 4, 5\}$ $\{6\}$

Utilizando el mismo método pero cambiando las medidas de disimilaridad entre empresas, se obtienen diferentes clasificaciones. Para ver como funcionan podemos comparar la forma en que se van agrupando los elementos a partir de las matrices definidas de dos maneras diferentes:



Como se puede ver, los elementos se van uniendo de una forma totalmente diferente. Por ejemplo, la empresa 1 en un caso se une inmediatamente con la 2 para formar el primer conglomerado, mientras que con la distancia de Mahalanobis son las últimas en agruparse.

3.5.2 Algoritmo de Ward

Trabajando con este algoritmo y las seis opciones posibles sobre la matriz de observaciones, luego de estandarizar las variables se obtienen las particiones en 2 y 3 conglomerados descriptas en los cuadros a y b.

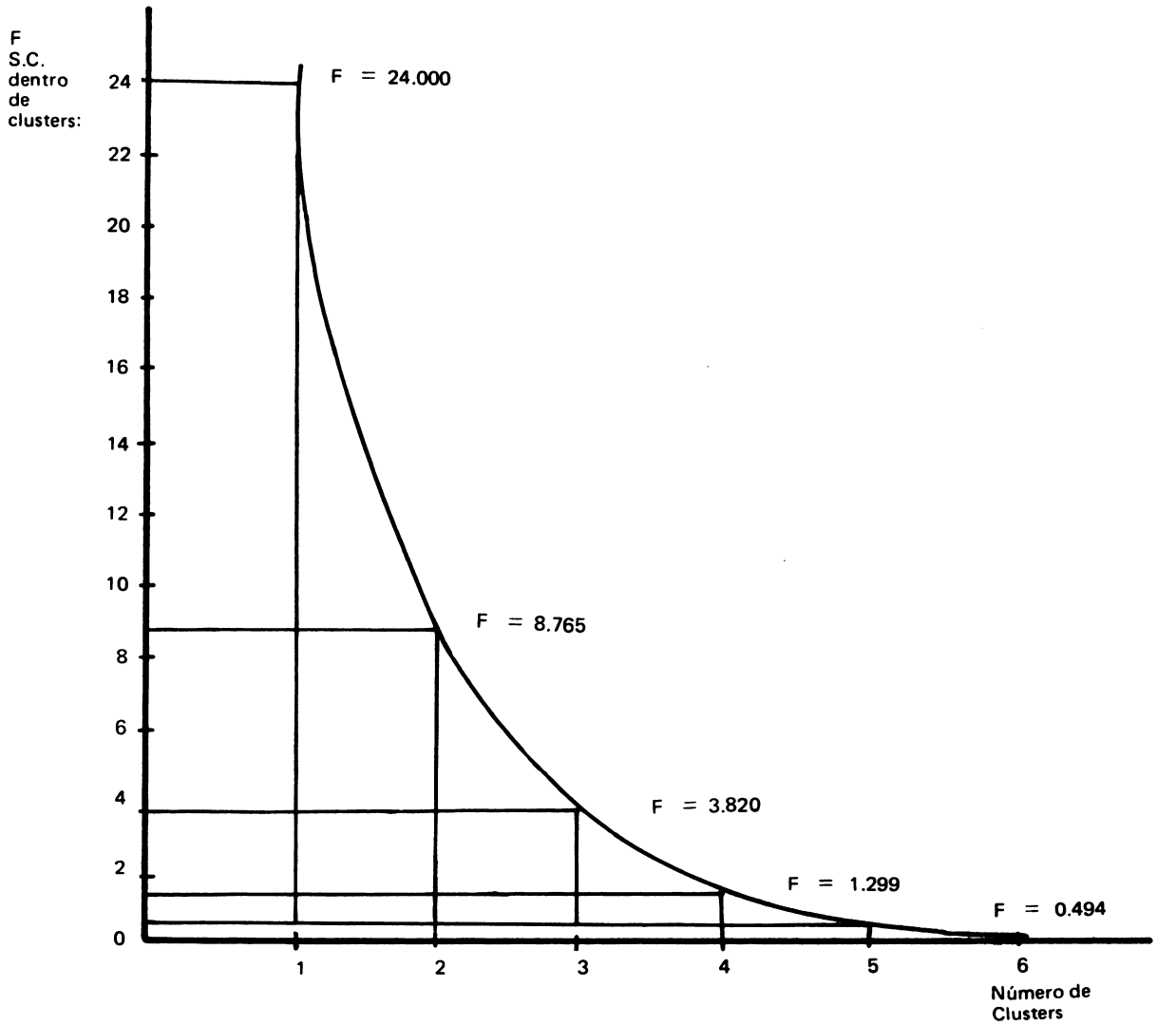
Para el algoritmo de Ward se obtiene además la suma de cuadrados dentro de clusters para cada iteración que aparece con el valor de la función objetivo F.

Como se trabajó con variables estandarizadas que tienen varianza unitaria, el valor que se obtiene al finalizar el análisis es de 24, que corresponde a la suma de cuadrados total ya que:

$$V\ (\text{total}) = \sum_{i=1}^{6} V(X_i) = 6$$

$$S.C. = V\ (\text{total}) \times n = 24$$

Podemos graficar el valor de F con respecto al número de clusters en la siguiente forma:



3.5.3 Algoritmo de Sparks

Al aplicar el algoritmo de Sparks sobre los datos estandarizados se obtuvo:

2 Clusters: {1, 2, 3} {4, 5, 6}

3 Clusters: {1, 2, 3} {4, 5} {6}

Para poder comparar con mayor facilidad las clasificaciones obtenidas, se representan en el cuadro siguiente los agrupamientos que resultaron de la aplicación de los distintos métodos de clasificación:

a) Partición en 2 conglomerados

METODO	CONGLOMERADOS	
	1	2
VAN RIJSBERGEN		
* Distancia euclideana	1, 2, 3, 6	4, 5
* Distancia euclideana estandarizada	1, 2, 3, 4, 5	6
* D. eucl. estandarizando y ponderando	1, 2	3, 4, 5, 6
* D. de Mahalonobis estandarizando	1	2, 3, 4 , 5, 6
VECINO MAS CERCANO	1, 2, 3 , 4 , 5	6
VECINO MAS LEJANO	1, 2, 3	4, 5, 6
MEDIANA	1, 2, 3	4, 5, 6
CENTROIDE	1, 2, 3	4, 5, 6
PROMEDIO DE GRUPO	1, 2, 3	4, 5, 6
WARD	1, 2, 3	4, 5, 6
SPARKS	1, 2, 3	4, 5, 6

b) Partición en 3 conglomerados

METODO	CONGLOMERADOS		
	1	2	3
VAN RIJSBERGEN			
* Distancia Euclideana	1, 2, 3	4, 5	6
* D. Euclideana estandarizando	1, 2, 3	4, 5	6
* D. Euclideana estandarizando y ponderando	1, 2	3, 4, 5	6
* D. de Mahalanobis estandarizando	1	2, 3, 4, 5	6
VECINO MAS CERCANO	1, 2, 3	4, 5	6
VECINO MAS LEJANO	1, 2, 3	4, 5	6
MEDIANA	1, 2, 3	4, 5	6
CENTROIDE	1, 2, 3	4, 5	6
PROMEDIO DE GRUPO	1, 2, 3	4, 5	6
WARD	1, 2, 3	4, 5	6
SPARKS	1, 2, 3	4, 5	6

En realidad, no se puede afirmar cuál o cuáles son las variables que están determinando una clasificación dada sin realizar previamente una dócima que lo justifique. Lo que se trató fue mostrar que aplicando los distintos métodos se puede llegar a resultados que pueden ser muy diferentes.

Cuando se trabajó con distancia euclideana ponderada se obtuvo una clasificación que intuitivamente, aparecía como determinada por la variable S.C./S.T. Podemos, entonces, plantear una tabla de contingencia para ver si existe una asociación entre el hecho de que las empresas pertenezcan a un cluster dado y el valor que toman para esa variable:

S.C./S.T.	Cluster 1	Cluster 2	f. m.
0 a 0.20	0 (1,33)	4 (2,66)	4
0.20 a 0.40	2 (0.66)	0 (1,22)	2
f . m .	2	4	6

$$T = \frac{(0 - 1{,}33)^2}{1{,}33} + \frac{(4 - 2{,}66)^2}{2{,}66} + \frac{(2 - 0{,}66)^2}{0{,}66} + \frac{(0 - 1{,}33)^2}{1{,}33}$$

$$= 1{,}33 + 0{,}66 + 1{,}33 + 1{,}33$$

$$\boxed{T = 4{,}65}$$

Nivel 5% ⟶ $X_1^2 = 3{,}84$

Nivel 1% ⟶ $X_1^2 = 6{,}63$

A nivel 5% se rechaza la hipótesis nula de ausencia de asociación entre las dos características, pero a nivel 1% se debe aceptar.

El estadígrafo de contingencia será:

$$C = \sqrt{\frac{4{,}65}{4{,}65 + 6}}$$

$$\boxed{C = 0{,}66}$$

máximo C = 0,707.

Se puede ver, en este caso, que la confianza con que podemos afirmar que los conglomerados obtenidos están relacionados con la variable S.C./S.T. no es muy grande, ya que a nivel 0,01 no se puede rechachar la hipótesis nula de ausencia de asociación. Sin embargo, el grado de asociación entre las dos características es muy alto, como lo indica el valor obtenido del estadígrafo de contingencia.

Al trabajar con distancia euclideana, sin estandarizar se puede pensar que la clasificación estaría determinada por la variable N.T. que está expresada en unidades más grandes. Podemos tratar de probar esto mediante la dócima de Kruskal y Wallis, para lo que debemos ordenar las empresas de acuerdo con el valor que presentan para la variable de interés.

Empresas	Ordenamiento
4	1a.
5	2a.
6	3a.
3	4a.
2	5a.
1	6a.

	CONGLOMERADOS	
	1	2
ORDENAMIENTO	3a. 4a. 5a. 6a.	1a. 2a.
R_j	18	3
R_j^2	324	9
n_j	4	2
R_j^2 / n_j	81	4, 5

$$H = \frac{12}{6\,(6+1)}\,(81 + 4{,}5) - 3(6+1)$$

$$H = 2{,}86$$

Nivel 10% $\quad X_1^2 = 2{,}71$

Nivel 5% $\quad X_1^2 = 3{,}84$

A nivel 10% se rechaza la hipótesis nula de ausencia de asociación entre las dos características, pero a nivel 5% se debe aceptar.

La confianza con que podemos afirmar que la clasificación obtenida está relacionada con la variable N.T. no es muy grande.

Análisis discriminante

Para aplicar el Análisis discriminante, como prueba de las clasificaciones obtenidas, se eligieron los grupos 1, 2, 3 y 4, 5, 6 que se formaron al aplicar el método de Ward y cinco de las seis alternativas de Ward.

Los resultados que se obtuvieron fueron:

* **Medias**

Grupo 1	:	23,33	0,16	0,31	381,67
Grupo 2	:	101,67	0,06	0,11	733,33

* **Distancia Generalizada de Mahalanobis**

$D^2 = 100,37$

$X^2 (4, 0.005) = 14,9$

* **Funciones discriminantes**

Función discriminante 1

Constante – 264,97

Coeficientes 3,29 1796,81 163,03 0,32

Función discriminante 2

Constante – 453,68

Coeficientes 4,33 2241,64 216,55 0,43

* **Evaluación de las funciones de clasificación para cada observación**

GRUPO	OBSERVACION	PROBABILIDAD ASOCIADA CON LA MAYOR FUNCION DISCRIMINANTE	FUNCION Nº
1	1 – (1)	1.00	1
	2 – (2)	1.00	1
	3 – (3)	1.00	1
2	1 – (4)	1.00	2
	2 – (5)	1.00	2
	3 – (6)	1.00	2

El valor de D^2 es altamente significativo; si se lo compara con el valor de tablas se ve que aún a nivel 0,5% se debe rechazar la hipótesis de que las clases pueden haber surgido al azar.

Las funciones discriminantes no son muy diferentes, pero se debe tener presente que se está trabajando con un número de observaciones muy reducido.

La clasificación de las empresas aparece como buena, en la medida en que todas las observaciones presentan una probabilidad de 1.00 de estar clasificadas correctamente. Sin embargo, en este caso no se podía esperar otro resultado debido a que se trabaja con 2 grupos con 3 observaciones cada uno y con un número relativamente alto de variables.

3.6 Referencias

1. ALONSO, A., "Productividad y Tipificación en la Agricultura de la República Oriental del Uruguay", Trabajos de Investigación Aplicada, CIENES (Santiago de Chile), 1975.

2. CEPAL, "Estudio sobre la clasificación económica y social de los países de América Latina", Documento de Información, E/CN. 12/878, 1971.

3. CONOVER, W.J., "Practical Nonparametric Statistics", J. Wiley, 1971.

4. EDWARDS, A. y CAVALLI-SFORZA, L., "A method for cluster analysis", Biometrics, vol. 21, No. 2: 362-375, 1965.

5. FLOREK, K. et al., "Taksonomia Wroclawska, Przege antrop., vol. 17: 193, 1951, Resumen en inglés.

6. GREEN, P., FRANK, R., y ROBINSON. P., "Cluster analysis in Test Market Selection", Management Science., Vol. 13, No. 8: 387-400, 1967.

7. HARRISON, I., "Cluster analysis", Metra, vol. 7, No. 3: 513-528, 1968.

8. I.I.C.A., "Seminario sobre métodos y problemas en tipificación de empresas agropecuarias", Vols. 1, 2 y 3, Serie de Informes, Cursos y Conferencias No. 92, Montevideo, 1975.

9. LANCE; G.N. and WILLIAMS, W.T., "A general theory of classificatory sorting strategies 1. Hierarchical Systems", Computer Journal, vol. 9: 373-80, 1967.

10. LEBART y FENELON, "Statistique et informatique apliquées", Dunod, 1973.

11. LING, R.F. y KILLOUGH, G.G., "Probability Tables for Cluster Analysis Based on a Theory of Random graphs", Journal of the American Stat. Association, Vol. 71, No. 354, 1976.

12. MACHADO, A., GALLO, J. y CAFFERA, J., "Manual de usuarios para programas de clasificación del IICA", IICA, Montevideo, 1976.

13. MORRISON, D.F., "Multivariate statistical methods", McGraw-Hill, 1967.

14. MORRISON, D. G., "Measurement problems in cluster analysis", Management Science, vol. 13, No. 12, 1967.

15. SPARKS, D.N., "Euclidean cluster analysis", Applied Statistics 22, 1973.

16. PRESS, S.J., "Applied Multivariate analysis", Holt-Rinehart-Winston, 1972.

17. PRETZER, D. and FINLEY, R., "Farm type classification, Another look at an old problem", American Journal Agric. Econ., vol. 56, No. 1, 1974.

18. VAN RIJSBERGEN, C.J., "A fast hierarchic clustering algorithm", Computer Journal, vol. 13: 324-326, 1970.

19. WISHART, D., "An algorithm for hierarchical classifications", Biometrics, 1969.

CAPITULO 4

Algunos comentarios sobre evaluación de clusterings.

Algunos comentarios sobre evaluación de clusterings.

4

Pedro Ferreira — CIENES - OEA

4.1 Introducción

El investigador que se propone analizar un conjunto de datos utilizando en determinada etapa del trabajo un cluster analysis (análisis de conglomeración) se enfrenta con una enorme gama de métodos, una difícil selección de atributos o variables, etc..

Es imposible dar en abstracto una estrategia de selección de variables y métodos que partiendo de un cierto conjunto de datos entregue como resultado un clustering ideal. El agrupamiento homogéneo de unidades depende de la definición de homogeneidad y ésta, a su vez, de los objetivos del trabajo en cuestión.

Sin embargo, es importante intentar detallar algunas características o propiedades óptimas que debería poseer una clasificación. Estas características pueden servir de guía en la selección de metodología y en los intentos de evaluación a posteriori de una clasificación.

4.2 Concordancia con determinado principio o criterio

Existen diferentes formas significativas de clasificar un conjunto de datos relevantes asociados a cierto problema, muchas de las cuales corresponden, o permiten descubrir al investigador, un determinado principio o criterio de agrupamiento inherente al problema estudiado.

La validez o utilidad de un clustering muchas veces es medida en términos de la capacidad de ejemplificar o poner en evidencia algún principio de agrupamiento. Un ejemplo de esta situación de búsqueda de concordancia está dado en Boussard y Petit (1966) quienes estudian un clustering de establecimientos agrícolas en cuanto a su dependencia del criterio "intensidad de la producción" o "intensidad de cultivo".

Cuando uno intenta comparar un clustering con una partición obtenida de acuerdo a cierto criterio, tres posibilidades diferentes deben ser consideradas:

4.2.1. **Dócimas o tests de independencia** – Estos tests se realizan formando una tabla de contingencia o de clasificación a dos criterios (el criterio del clustering y el externo). Una discusión de procedimientos alternativos para probar la independencia puede encontrarse por ejemplo en Kendall y Stuart (1967, cap. 33). Tests para muestras grandes suelen basarse en la estadística

$$\chi^2 = n \sum_{ij} [n_{ij}^2 / n_{i.}\ n_{.j}] - 1 \quad \text{donde} \quad n_{i.} = \sum_j n_{ij} \ ; \ n_{.j} = \sum_i n_{ij}$$

la cual tiene distribución χ^2 con $(r-1)(c-1)$ grados de libertad.

4.2.2. **Medidas de dependencia** – Se intenta describir la magnitud de la dependencia de una clasificación respecto de cierto criterio.

Por ejemplo podría medirse la dependencia de una clasificación de establecimientos agrícolas respecto de los tipos de suelos o de los tamaños de los establecimientos.

Son útiles para tal propósito las medidas usuales de dependencia en una tabla de contingencia; una discusión puede encontrarse en Goodman y Kruskal (1954).

Una buena medida frecuentemente utilizada viene dada por la información para discriminación o ganancia de información.

$$I = \sum_{i,j} \frac{n_{ij}}{n} \log_2 \frac{n\ n_{ij}}{n_{i.}\ n_{.j}}$$

Esta medida es mayor que cero salvo en el caso en que

$$n_{ij} = \frac{n_{i.}\ n_{.j}}{n}$$

4.2.3. **Medidas de congruencia** – En los casos anteriores estábamos interesados en cuantificar en qué medida una clasificación estaba influenciada o dependía de ciertos factores. En este caso nos interesa medir la concordancia o perfecto ajuste con la partición inducida por otro criterio.

Una medida adecuada viene dada a través de la diferencia simétrica entre las relaciones correspondientes a las particiones dadas por ambos criterios (ver, por ejemplo, Jardine y Sibson (1971, sección 11.3)).

4.3 Estabilidad de una clasificación

Sin embargo el investigador debe diferenciar aquellas situaciones en que un método revela ó confirma una interesante estructura inherente a ciertos datos, o a cierto problema, y aquellas en que la metodología impone una estructura inexistente.

Estas consideraciones están ligadas a la estabilidad de las clasificaciones frente a diferentes selecciones de un conjunto razonablemente amplio de atributos o variables tomados del conjunto general de atributos elegidos para el problema en cuestión.

Una clasificación inestable no es útil para hacer predicciones basándose en ella, puesto que no es razonable pensar que nuevos atributos se comporten homogéneamente dentro de los grupos formados.

Estos nuevos atributos podrían, en el problema que nos preocupa, ser índices cuantificadores del resultado de la aplicación de ciertas políticas económicas a los diferentes conglomerados. En una clasificación inestable no podemos esperar que estos índices se comporten homogéneamente dentro de los clusters.

Se hace necesario entonces hacer un estudio de la estabilidad de una clasificación y de disponer de criterios de comparación de clasificaciones provenientes de diferentes selecciones de atributos o variables.

Otro aspecto de la estabilidad se relaciona a la sensibilidad de un método frente a la carencia de una información completa, o de conjuntos de unidades que no están adecuadamente representados en los datos.

El investigador está interesado en poseer una medida del grado de concordancia entre la clasificación que él obtiene y la que se obtendría si se tuviera un conjunto más amplio de datos.

4.4 Costos y capacidades requeridas en la tarea computacional

Es muy excepcional que en un problema de clustering no sea necesario el uso de un computador y es entonces un criterio importante de selección y comparación de metodologías los requerimientos de memoria y tiempo de procesamiento de los diversos métodos.

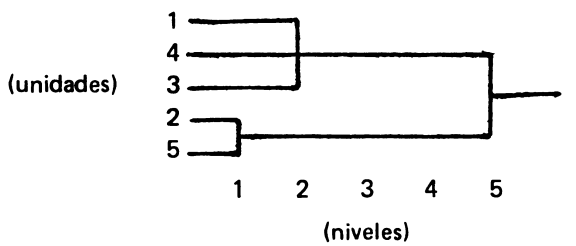
En general, los métodos llamados "de agrupamiento al centroide más próximo" (por ejemplo: Sparks toma K puntos y agrupa alrededor de ellos, calcula sus centroides y vuelve a agrupar, ver Anderberg (1973, Cap. 7) son más económicos que los métodos jerárquicos debido a que en ellos no es necesario ir calculando y guardando en memoria una matriz de similaridades. No es tampoco necesario almacenar el conjunto de datos, los cuales pueden leerse secuencialmente de una cinta o disco magnético. Esta ventaja se ve a veces disminuida por el hecho de que algunos de estos métodos requieren que el analista determine en forma exógena el número de clusters o un intervalo que contenga tal número.

Un dato interesante sobre los métodos jerárquicos en que el tiempo de procesamiento y la memoria requerida crecen en forma aproximadamente proporcional al cuadrado del número de entidades. Existen sin embargo algunas estrategias que permiten particionar el análisis de grandes conjuntos de datos en bloques de tamaño manejable, a los cuales se les aplican procedimientos de clustering agrupando sus unidades a ciertos niveles (bajos) iniciales de similaridad. Luego estos clusters, que surgen de los diferentes bloques de datos, son en una segunda etapa conglomerados a su vez. Una descripción ejemplificada fue presentada por Anderberg (1973, p. 185 - 187).

4.5 Comparación de clasificaciones jerárquicas

Un método que ha sido muy difundido para evaluar la bondad de una clasificación jerárquica, consiste en calcular la correlación entre los elementos de la matriz de similaridades original y la matriz de similaridades derivada de un árbol jerárquico. Este coeficiente ha sido llamado "de correlación cofenotípica", ver Farris (1969) y Sokal y Rohlf (1962).

Un ejemplo de su cálculo se presenta a continuación. Supongamos que un clustering jerárquico de cinco entidades ha llevado al siguiente árbol:



Las unidades 2, 5 se unieron formando un cluster a un nivel jerárquico 1; las unidades 1, 4, 3 se unen dormando otro grupo a nivel 2 y finalmente a nivel 5 se unen los dos grupos.

A partir de este árbol se construye una matriz de similaridades derivada

1					
2	5				
3	2	5			
4	2	5	2		
5	5	1	5	5	
	1	2	3	4	5

y luego se calculará la correlación entre los 10 valores en dicha matriz y los diez valores de la matriz original.

Este esquema puede usarse para comparar diferentes clasificaciones de un mismo conjunto de datos, calculando para cada una de ellas la matriz de similaridades derivada y luego los coeficientes de correlación entre pares de matrices derivadas; ver Borko (1968, p. 15-28). Un problema asociado a este método consiste en que los datos a manejar en las matrices derivadas son muchas veces de tipo ordinal, correspondiendo al valor numérico de la iteración en que cada par fue unido. Es mejor para esta situación usar coeficientes de asociación basados en órdenes. Una posibilidad consiste en recurrir al coeficiente tau de Kendall, aunque éste presenta la dificultad de que si varios individuos tienen el mismo rango, su valor no tiene una interpretación simple. Por tal motivo, se recomienda el uso del índice gamma de Goodman y Kruskal (1954).

Otros métodos de comparación de matrices de disimilaridad (o de similaridad) pueden lograrse eligiendo diferentes métricas en un espacio Euclideano y considerando una matriz de similaridades como un punto en $R^{n\,(n-1)/2}$

Por ejemplo, Jardine y Sibson (1971, p. 103-107) consideran :

(i) la métrica usual

$$D\,(d_1, d_2) = \left(\Sigma\,(d_1\,(A, B) - d_2\,(A,B)\ )^2\ \right)^{1/2}$$

ii) la de la máxima diferencia entre pares de coordenadas

$$D\ (d_1, d_2) = \max \left|\ d_1\,(A, B) - d_2\,(A, B)\ \right|$$

(iii) la suma de distancias entre coordenadas

$$D\,(d_1, d_2) = \Sigma \left|\ d_1\,(A, B) - d_2\,(A, B)\ \right|$$

donde $d_1(A,B)$ representa la disimilaridad entre dos objetos A, B en la primer matriz y $d_2(A,B)$ lo mismo respecto de la segunda.

Otra medida de este tipo puede hallarse en Hartigan (1967), quien compara dos matrices de similaridad mediante la fórmula :

$$D_H (S_1, S_2) = \Sigma\ W(i, j)\ (S_1 (i, j) - S_2 (i, j))^2$$

donde S(i,j) representa la similaridad entre los objetos i, j y W(i,j) es una ponderación. D_H es entonces una distancia Euclideana ponderada.

Hartigan estudia un método de clustering que minimiza D_H cuando ésta es calculada entre la matriz original y la derivada del árbol jerárquico.

4.6 Evaluación de la similaridad entre particiones

4.6.1. Un método de comparación de dos particiones surge de la construcción de la siguiente tabla de contingencia:

Partición I		**Partición II**		
Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 3
1	2	1	2	4
3	4	3	9	5
9	5	8	12	6
11	6	10	13	7
15	7	15	14	11
	8			
	10			
	12			
	13			
	14			

Tabla de Contingencia

Partición I / Partición II	1	2	3
1	3	1	1
2	2	4	4

Esta tabla fue usada por Borko (1968, p. 62-80) tomando a partir de ella una medida de asociación como ser la estadística χ^2 o la contingencia media ϕ dada por

$$\phi^2 = \chi^2 / n\ .\ .$$

4.6.2. Otro enfoque surge de considerar las unidades (o establecimientos agrícolas) de a pares y de registrar como "similares" aquellas unidades que forman un par perteneciente a un mismo cluster y como "disímiles" aquellas en diferentes clusters. Entonces si se considera el cociente entre el número de pares similares entre el total de comparaciones de a pares ($\frac{1}{2}$ N (N – 1)) se tiene la medida de similaridad entre particiones de Rand (1971). El complemento a uno de esta medida fue usado también por Green y Rao (1969, p. 363).

4.7 Evaluación de las influencias de las diferentes variables en una clasificación jerárquica.

Una forma de medir la influencia de cada variable en una clasificación jerárquica, se logra examinando el comportamiento de la suma de cuadrados dentro de grupos a lo largo del proceso de conglomeración.

Se suele calcular para cada variable y para cada etapa de la clasificación jerárquica la llamada proporción no explicada de la varianza, como el cociente entre la suma de cuadrados dentro y la suma de cuadrados total (alrededor de la media).

Se distingue entonces entre aquellas variables cuyas proporciones no explicadas son grandes desde las primeras etapas de la conglomeración y aquellas cuyas proporciones permanecen pequeñas durante casi todo el proceso, creciendo solamente en las últimas etapas del clustering. A las primeras les llamaremos variables latentes o aletargadas y a las segundas variables dominantes.

Es de esperar que la eliminación de variables aletargadas no influencie el resultado del clustering debido a que obviamente el criterio solamente se estaría preocupando de homogeneizar respecto de las variables dominantes.

Esta clase de análisis puede ser útil para la selección del conjunto de variables a ser utilizadas o para la ponderación de las mismas.

Un programa en lenguaje Fortran que realiza este análisis puede encontrarse en Anderberg (1973, p. 340-42).

En un clustering no jerárquico puede resultar interesante el analizar variable a variable la proporción no explicada de la varianza para comparar los diversos niveles de homogeneidad logrados.

4.8 Referencias

ANDERBERG, M. R., Cluster Analysis for Applications, Academic Press (1973).

BOUSSARD, J. M. y PETIT, M., Problemes de l'accession a l'irrigation, Ins. Nat. de la Recherche Agron, (1966).

BORKO, H., BLANKENSHIP, D. A. y BURKET, R. C., On line information retrieval using associative indexing. RADC – TR – 68 – 100, AD 670195. Systems Develop. Corp., Santa Monica, California.

FARRIS, J. S., On the cophenic correlation coefficient, Syst. Zool. 18, No. 3, 279-285 (1969).

GOODMAN, L. A. y KRUSKAL, W. H., Measure of association for cross classifications, Jour. Amer. Stat, Ass. 49, 732 - 764 (1954).

GREEN, P. E. y RAO, V. R., A note on proximity measures and cluster analysis Jour. Marketing Res. 6, 359-364 (1969)

HARTIGAN, J. A., Representation of similarity matrices by trees, Jour, Amer. Stst. Ass. 62, 1140-1158 (1967).

JARDINE, N. and SIBSON, R. Mathematical Taxonomy, Wiley, New York (1971).

KENDALL, M. G. and STUART, A. The Advanced Theory of Statistics, Vol. II, Inference and Relationship, Griffin, London (1967).

LERMAN, I. C., Les Bases de la Classification Automatique, Gauthier-Villars, Paris (1970)

RAND, W. M. Objetive criteria for the evaluation of clustering methods. Jour. Amer. Stat. Ass. 66, 848-850 (1971).

SOKAL, R. R., and ROHLF, F. J. The comparison of dendrograms by objective methods, Taxon 11, 33-40 (1962).

ZAHN, C. T., Approximating symetric relations by equivalence relations, SIAM J. Appl. Math. 12, 840-847, (1964).

Apéndice

Algunos desarrollos recientes en evaluación de clusterings: Métodos basados en la teoría de grafos aleatorios

Existe muy poca teoría, hasta la fecha, que permita a un analista que ha utilizado un determinado método de cluster analysis decidir si sus clusters corresponden a una estructura real o si son completamente irrelevantes. Desde este punto de vista podría afirmarse que el cluster analysis ha proliferado fundamentalmente como una especie de arte, pero aún permanece en una etapa primitiva como rama de la ciencia.(1)

En el caso de un clustering jerárquico, el método de evaluación más comunmente usado consiste en el cálculo de algún coeficiente de concordancia entre la matriz de similaridades original y la derivada de los niveles de similaridad en que un par de objetos son unidos en un mismo cluster por primera vez. Esta concordancia es entonces medida a través de índices tales como el gamma de Goodman-Kruskal o el tau de Kendall.

Es evidente que para comparar tales índices o para poder decir desde un punto de vista estadístico cuan grande es un determinado valor observado, es necesario tener alguna noción acerca de la distribución de probabilidades de tales índices.

Por otra parte, las distribuciones de probabilidad a estudiar dependen del método elegido y es entonces difícil pensar en una solución general del problema.

Recientemente, partiendo de la teoría de distribuciones asociadas a grafos aleatorios, diferentes autores han presentado métodos que permiten diferenciar entre agrupamientos que pueden considerarse aleatorios y aquellos que no lo son.

La forma de asociar un grafo a un clustering, consiste en hacer corresponder a cada objeto un vértice, o nodo, o punto, en un grafo, y cuando la disimilaridad entre un par de objetos sea menor que un cierto nivel fijo, h, unimos los correspondientes vértices del grafo con una línea, o lado.

En este tipo de representación, un cluster del tipo single-link corresponde a un subgrafo, o componente conexa, o conectada (cada par de nodos puede ser unida por una sucesión de lados, o líneas, continua).

Consideremos por un momento grafos con m líneas o lados. Dado un conjunto de n nodos o vértices, existen C_2^n posibles lados y, si suponemos que se eligen al azar con igual probabilidad, m de ese total de C_2^n lados, tendremos un total de :

$$\left(\binom{n}{2} \right)_m$$

posibles grafos igualmente probables. Un grafo aleatorio es simplemente un elemento tomado al azar de este conjunto.

La idea es comparar un clustering logrado por un investigador con la distribución aleatoria de los grafos correspondientes para determinar la probabilidad de ocurrencia aleatoria de una estructura de ese tipo.

Un criterio de comparación usado por Ling y Killough[1], considera el mínimo número de ejes en que un grafo se vuelve conectado o conexo. Para el método de single-link, ese trabajo desarrolla una tabla de la distribución de probabilidades de esa variable en grafos aleatorios. Esta tabla es luego usada en casos específicos de clasificación para determinar el grado de aleatoriedad de los resultados obtenidos.

Otro criterio de comparación usado por los autores recientemente mencionados, es el del valor esperado de "componentes" en un grafo aleatorio con m lados. Una componente es definida como un subgrafo conectado aislado, o sea un cluster tipo single-link o un punto aislado. Este criterio también es utilizado en la comparación de clusters de clasificaciones específicas para determinar su grado de aleatoriedad.

Un approach similar al recientemente mencionado fue presentado por Baker y Hubert[2] para determinar el grado de aleatoriedad de un clustering de tipo complete-link.

Referencias del apéndice

1. LING, R. F. y KILLOUGH, G. G., Probability tables for cluster analysis based on a theory of random graphs, Jour. Amer. Stat. Ass. 71, No. 354, June 1976.

2. BAKER, F. B. y HUBERT, L. J., A graph-theoretic approach to goodness-of-fit in complete-link hierarchical clustering, Jour. Amer. Stat. Ass. , 71, No. 356, dec. 1976.

CAPITULO 5

Comentarios sobre procesos de tipificación y su validación.

Comentarios sobre procesos de tipificación y su validación.

5

Mario Kaminsky — CIENES - OEA

5.1 Resumen general

En este trabajo se presentan diversos comentarios sobre la praxis de procesos de tipificación y su validación. Dado que ya se ha avanzado lo suficiente en el área de aplicaciones de diversas técnicas y metodologías de tipificación propiamente dicha, parece llegado el momento de ocuparse del siguiente paso natural: los procesos de validación de tipificaciones logradas o que se estén intentando o se intenten en el futuro. Ello es lo que determina el énfasis de estos comentarios sobre validación.

Reconocidamente, el objetivo de los comentarios es modesto: iniciar la discusión sobre tan importante tema, y resaltar algunas preocupaciones y advertencias que surgen de dichas preocupaciones, en relación con el área que se intenta enfatizar, así como ilustrar someramente algunos conceptos básicos y temas relacionados.

Luego de una Introducción, los comentarios se dirigen a las siguientes áreas generales: tipificación, validación e identificación; antecedentes sobre validación y algunas aplicaciones e ilustraciones; validación de metodologías de validación; cómo no validar, vía "concordancia"; criterios de estabilidad en procesos de validación; evaluación de similaridades entre participaciones o grupos; (una ilustración de) grafos aleatorios; y el empleo de correlación canónica en procesos de validación. Seguidamente se incluyen las referencias citadas en los comentarios, y dos anexos: uno sobre la traducción de un trabajo al que profusamente se hace referencia en el texto, el otro siendo la tabla estadística de la distribución de Chi-Cuadrado.

5.2 Introducción

En la primera reunión de estos Seminarios sobre Métodos y Problemas en Tipificación de Empresas Agropecuarias el énfasis general fue puesto sobre la comparación, y consecuente evaluación, de métodos "tradicionales" encontrados en la praxis de la tipificación de empresas agropecuarias y métodos estadísticos y cuasi-estadísticos disponibles pero no usualmente empleados. En tal sentido existió acuerdo acerca de que existe "poca interacción entre las técnicas estadístico - informáticas y los métodos y teorías de las ciencias sociales". También acerca de que "Un mejoramiento en la situación comentada debe surgir de la aplicación consciente y metódica de estas técnicas (estadísticas), lo que debiera estar precedido por razonables actividades de aprendizaje y entrenamiento teórico y aplicado".

Entre las críticas a métodos "tradicionales" se señaló en la misma ocasión que "La clasificación por tramos en contraposición con el uso de clustering, aparece intuitivamente como poco natural y eficiente. Esta crítica se debe a que las clases así conformadas están restringidas a ser rectángulos, pese a que los tramos de una variable pueden no ser óptimos para ciertos tramos de otra/s variable/s de cruzamiento"; señalándose a continuación posibles vías de mejoramiento de la situación. Se hizo notar también el estado poco satisfactorio en que se encuentra la cobertura de áreas como las de "Evaluación y selección de las técnicas de tipificación que se adecúen a los objetivos y (que) resulten operativas".

Es de esperarse que, entre lo que va desde entonces a hoy, se haya avanzado en dicho camino, aunque indudablemente debe quedar mucho por recorrer.

En la misma primera reunión se prestó alguna atención, aunque tangencial, a otra área problemática, más específica. Continuando con la cita de las Conclusiones y Recomendaciones a que entonces se llegó, puede recordarse en relación con este punto, lo siguiente:

> ". . . una cuidadosa selección de atributos para tipificar presupone un análisis consciente del sistema que se analiza, incluyendo tareas de validación objetiva del proceso".
>
> "Las técnicas en consideración aportan reglas de decisión para juzgar la bondad o validez de clasificaciones ensayadas".
>
> ". . . las técnicas de análisis a posteriori . . . pueden ayudar para la interpretación del funcionamiento de las empresas tipo, por ejemplo: en el análisis causa - efecto de variables de estas empresas".
>
> ". . . la combinación de diversos métodos y técnicas puede ser conveniente y necesaria. En particular, el análisis discriminante podría emplearse con soltura combinado con operaciones de muestreo que en forma económica permitan, a través de estudios exhaustivos, conocer con un grado razonable de aproximación los atributos clasificatorios generales de interés".
>
> "La obtención de agrupamientos similares mediante el uso de diversas técnicas, sugeriría la robustez del proceso tipificatorio que se intenta. La cuantificación de la homogeneidad intra - grupos obtenida también permitiría comparar técnicas".
>
> "Una posible enunciación de (los pasos recomendables) es la siguiente: . . . 5. Implementación de la metodología. Caracterización de empresas tipo y determinación de su representatividad en términos del universo a estudiar . . ."

Lo anterior apunta entonces a un área problemática que se podría denominar "validación de procesos de tipificación". Es de esperarse que en esta segunda reunión ella reciba un tratamiento más profundo, en concordancia con el avance logrado en el área que se enfatizó en la primera.

Es el propósito de estos Comentarios tender a lograr dicho objetivo. De entre los pocos trabajos conocidos con antelación a esta segunda reunión, y a ser presentados en ella, sin duda apuntan en el mismo sentido los de Pedro Ferreira (6) y Alfredo Alonso (2), este último en su sección 3.4.

5.3 Sobre tipificación, validación e identificación

Por qué es importante la validación de procesos de tipificación? Implícita o explícitamente todo proceso de tipificación tiene por objeto la identificación de "tipos". A su vez la identificación de "tipos" tiene por objetivo la identificación de estructuras. Y por qué es importante la identificación de estructuras? Porque es el conocimiento de los parámetros estructurales, vs. el concerniente a los parámetros de "forma reducida" de un modelo, el único que de hecho permite evaluar las bondades relativas de "políticas estructurales" alternativas, ingrediente necesario de toda acción orientada a actuar sobre el medio, para adecuarlo a

objetivos deseados, adecuados a su vez a la optimización de funciones de bienestar o "ganancia" relevantes. El uso de políticas rutinarias (no estructurales) exhibe una muy limitada capacidad de cambio económico-social*. Pero debe tenerse en cuenta que una poco apta o errónea identificación de estructuras podría tener efectos aún más nocivos que la propia ignorancia de los parámetros estructurales relevantes; en otras palabras, podría llegar a ser peor el (mal) remedio que la enfermedad. De allí que, siendo la tipificación un medio para la identificación de estructuras, podría llegar a ser peor una (pobre/débil/errónea) tipificación que una "no tipificación". Aquí radica la importancia de la Validación.

Ello sugiere o, mejor aún, impone, que vis-à-vis todo proceso de tipificación, se diseñe o implemente un adecuado proceso de validación de los resultados del primero. En otras (de Pedro Ferreira) palabras: ". . . el investigador debe diferenciar aquellas situaciones en que un método revela o confirma una interesante estructura inherente a ciertos datos, o a cierto problema, y aquéllas en que la metodología impone una estructura inexistente"**.

5.4 Sobre antecedentes sobre validación y algunas aplicaciones e ilustraciones

En la Introducción del Editor a la publicación que resultó de la primera reunión de estos seminarios se puntualizaba que "La caracterización de "tipos" y "casos típicos" se hace usualmente en base a apriorismos intuitivos. Estos apriorismos pueden tener mayor o menor grado de fundamentación teórica y suelen emplear la opinión de "conocedores calificados"***.

En aquella ocasión, el autor de estos Comentarios comentó (11) un trabajo (3) que respondía a dicha descripción. Los tipos de empresa se construían por cruzamiento de "los estratos de superficie total de la unidad de producción con aquellos otros atributos que –A JUICIO DEL EQUIPO REGIONAL– permitan discriminar mejor los sistemas. . .". Pero ya aquí la idea de validación se hacía presente, a través de conceptos tales como "representatividad", descripción preliminar "con una base tan objetiva como sea posible", "caracterización definitiva de los tipos de empresa", etc. De hecho, una subsección completa**** se dedicaba al "Análisis de la Representatividad de los Tipos de Empresas y Verificación de la **Validez** de la Caracterización Realizada". Más específicamente, la idea consistía en la incorporación de **elementos objetivos** que permitieran la **cuantificación** de los tipos de empresa, mediante una sistematización que le diera **"validez estadística"** a la información básica. La misma preocupación estaba presente para la validación de la información surgida de los estudios de caso que se preveían para cada "empresa tipo" (representativa de un "tipo de empresa").

Sin embargo, algunas de las aplicaciones e ilustraciones incluídas en (8), la sección correspondiente de (7)***** y el trabajo arriba aludido, fueron los únicos que mostraron una explícita preocupación acerca de los procesos de validación de las tipificaciones propuestas o ensayadas. A su vez, sin embargo, el desnivel entre lo metodológico y lo operativo, advertido en el correspondiente Comentario al trabajo aludido, resultaba en la ausencia de criterios, métodos y técnicas explícitas que permitieran la implementación de "un intenso objetivo de validación" al cual someter "las tipificaciones en proceso con un esquema tan subjetivo". Se comentaba también que el proceso de "ajuste hacia arriba" que era de esperarse, conduciría a un deseable mayor uso de dócimas o pruebas estadísticas en el proceso al que dicho proyecto se refería.

---

* Sobre el particular, vid. (9) y (10). En el contexto de un problema y aplicación específicos, cf. también (8), pp. 3 - 7.

** (6), p. 3.

*** (5), p. 1.

**** (3), p. 35 (mayúsculas del autor de estos Comentarios).

***** (7), subsección II, pág. 20.

A continuación se hará referencia sucinta a las partes del documento (8) en que se presta atención directa o indirecta a problemas de validación y técnicas asociadas, siguiendo el orden de presentación contenido en el mismo:

Sección II. INTRODUCCION, p. 2, tercer párrafo.

Sección III. APLICACIONES, subsección A. TIPIFICACION DE EMPRESAS AGROPECUARIAS TAMBO SEGUN CAPACIDAD EMPRESARIAL, pp. 8-10.

subsección B. TIPIFICACION EN AGRICULTURA – ANALISIS DE CONGLOMERACION, KRUSKAL Y WALLIS Y TABLAS DE CONTINGENCIA, Resultados de Análisis de Kruskal y Wallis y Tablas de Contingencia, pp. 29-34.

Sección IV. ILUSTRACIONES, subsección A. TIPIFICACIONES DE EMPRESAS AGROPECUARIAS TAMBO DE LA CUENCA LECHERA DEL CENTRO SANTAFECINO (ARGENTINA) – HISTOGRAMAS, TABLAS DE CONTINGENCIA, ANALISIS DISCRIMINANTE, DE COMPONENTES PRINCIPALES Y DE CONGLOMERACION.

Resultados de Histogramas, pp. 43-47.
Resultados de Tablas de Contingencia, p. 48.
Resultados del Análisis Discriminante, pp. 48-53.
Resultados de Análisis de Conglomeración, p. 67.
("índice de clasificabilidad")

subsección B. TIPIFICACIONES DE ESTADOS DEL BRASIL POR INDICADORES ECONOMICOS GENERALES Y COMERCIO INTERESTADUAL POR VIAS INTERNAS – ANALISIS DE COMPONENTES PRINCIPALES, DE CONGLOMERACION Y DISCRIMINANTE.

Resultados del Análisis Discriminante, pp. 72-73.

APENDICE 1. EL METODO DE ANALISIS DISCRIMINANTE EMPLEADO EN LA APLICACION DE SECCION III.A, E ILUSTRACIONES DE SECCIONES IV.A y IV.B, p. 82-83.

APENDICE 4. LAS TECNICAS DE TABLAS DE CONTINGENCIA Y DE KRUSKAL Y WALLIS EMPLEADAS EN LAS APLICACIONES DE LA SECCION III.C E ILUSTRACIONES DE LA SECCION IV.A, pp. 105-129.
(Incluye Correlación por Orden de Rangos, p. 122-129).

APENDICE 5. DATOS Y RESULTADOS COMPLEMENTARIOS DE LAS APLICACIONES INCLUIDAS EN SECCION III.C, pp. 133-134.

APENDICE 7. DATOS Y RESULTADOS COMPLEMENTARIOS DE LAS ILUSTRACIONES INCLUIDAS EN SECCION IV.A. RESULTADOS DE TABLAS DE CONTINGENCIA, pp. 140-142.

APENDICE 9. NOTAS Y CITAS, (12), p. 148.

5.5 Sobre validación de metodologías de validación

No hay nada mágico en la Estadística que permita usarla para sustituir una adecuada metodología, una teoría adecuada, y también por qué no decirlo, el sentido común y la lógica. Las validaciones estadísticas deben ser empleadas con cuidado, para que sus aplicaciones se difundan, y en última instancia, sean realmente útiles*. La ausencia de validación positiva de un proceso de tipificación puede deberse a defectos en el diseño de la metodología de validación. Esto último puede incluir defectos en la teoría o conjunto de hipótesis sobre los que se fundamenta una tal metodología de validación. Pero las técnicas de validación disponibles son capaces también de detectar este último tipo de situaciones. Así, en un reciente trabajo cuyo objetivo no era propiamente de tipificación (1), se consideraba que, sobre la base de la evidencia empírica incluída en el Cuadro No. 1 reproducida más abajo, existía una asociación de tipo negativo o inverso entre "desarrollo" relativo (medido por nivel de ingreso per cápita) y desigualdad en la distribución de dicho ingreso (medido por la participación del 40% inferior):

> "Los **países desarrollados** se distribuyen en forma pareja entre las categorías de baja y moderada desigualdad. . . . La mayor parte de los **países subdesarrollados** presenta una desigualdad relativa acentuadamente más grande que los países desarrollados" **.

Sin embargo, un simple análisis de Tabla de Contingencia, construído sobre la base de esa misma evidencia empírica por el autor de estos Comentarios, **no permitía rechazar,** a los niveles usuales de significación estadística, la hipótesis de INDEPENDENCIA entre ambos atributos. Los respectivos resultados son como se indica en el Cuadro No. 1.

Cuadro 1

Tabla de Contingencia

Desig. / Nivel	Alta	Mediana	Baja	Total
Bajo	12(9.85)	6(7.88)	8(8.27)	26
Mediano	10(7.95)	6(6.36)	5(6.68)	21
Alto	3(7.20)	8(5.76)	8(6.05)	19
Total	**25**	**20**	**21**	**66**

donde los estadísticos entre paréntesis () indican las respectivas "frecuencias teóricas".

---

* Esto apunta en el mismo sentido de la advertencia incluída entre las Conclusiones y Recomendaciones del Seminario (5), sobre la base de "la consideración de que cualquier aplicación mecanicista y torpe de las técnicas lleva a condiciones nocivas desde un doble punto de vista:

- los resultados son inconducentes para los objetivos explicitados
- se derivan rechazos de las técnicas por usuarios potenciales que sí podrían obtener resultados conducentes.

Estos efectos nocivos dañan a las técnicas estadísticas y a la estadística en general, pero más dañan en el sentido de que no se obtienen resultados útiles que podrían ser obtenidos".

** (1), p. 3

Cálculo del estadístico Chi-Cuadrado:

$$\chi^2 = 4.6225/9.85 + 8.5344/7.88 + .0729/8.27 + 4.2025/7.95 + .1296/6.36 +$$

$$+ 2.8224/6.68 + 17.64/7.20 + 5.0176/5.76 + 3.8025/6.05 =$$

$$= .496 + .449 + .009 + .529 + .020 + .423 + 2.450 + .871 + .629 =$$

$$\chi^2 = 5.849,$$

que, contrastado con los valores tabulados

$$\chi^2_{4GL,\ .010} = 13.277 \ ; \ \chi^2_{4GL,\ .025} = 11.143 \ ; \ \chi^2_{4GL,\ .05} = 9.488,$$

no permite rechazar la hipótesis nula de independencia.

Cuadro 2

Doble clasificación de los países por nivel de ingreso y por desigualdad

ALTA DESIGUALDAD Participación del 40% inferior menos del 12%					MODERADA DESIGUALDAD Participación del 40% inferior entre el 12% y el 17%					POCA DESIGUALDAD Participación del 40% inferior 17% y más					
País (año)	PNB per cápita US$	40% inferior	40% medio	20% superior	País (año)	PNB per cápita US$	40% inferior	40% medio	40% superior	País (año)	PNB per cápita US$	40% inferior	40% medio	40% superior	
Kenia (1969)	136	10.0	22.0	68.0	Birmania (1958)	82	16.5	38.7	44.8	Chad (1858)	78	18.0	39.0	43.0	
Sierra Leona (1968)	159	9.6	22.4	68.0	Dehomey (1959)	87	15.5	34.5	50.0	Sri Lanka (1969)	95	17.0	37.0	46.0	
Irak (1966)	200	6.8	25.2	68.0	Tanzania (1967)	89	13.0	26.0	61.0	Niger (1960)	97	18.0	18.0	42.0	
Filipinas (1971)	239	11.6	34.6	68.0	India (1964)	99	16.0	32.0	52.0	Paquistán (1964)	100	17.5	37.5	45.0	
Senegal (1960)	245	10.0	26.0	64.0	Madagascar (1960)	120	13.5	25.5	51.0	Uganda (1970)	126	17.1	35.8	47.1	*
Costa de Marfil (1970)	247	10.8	32.1	57.1	Zambia (1959)	230	14.5	28.5	57.0	Tailandia (1970)	180	17.0	37.5	45.5	
Rhodesia (1968)	252	8.2	22.8	69.0						Corea (1970)	235	18.0	37.0	45.0	
Túnez (1970)	255	11.4	33.6	55.0						Taiwan (1964)	241	20.4	39.5	40.1	
Honduras (1968)	265	6.5	28.5	65.0											
Eduador (1970)	277	6.5	20.0	73.5											
El Salvador (1969)	295	11.2	36.4	52.4											
Turquía (1968)	282	9.3	29.9	60.8											
Malasia (1970)	330	11.6	32.4	56.0	República Dominicana (1969)	323	12.2	30.3	57.5	Surinam (1962)	394	21.7	35.7	42.6	
Colombia (1970)	358	9.0	30.0	61.0	Irán (1968)	332	12.5	33.0	54.5	Grecia (1957)	500	21.0	29.5	49.5	
Brasil (1970)	390	10.0	28.4	61.5	Guyana (1959)	550	14.0	40.3	45.7	Yugoslavia (1968)	529	18.5	40.0	41.5	
Perú (1970)	480	6.5	33.5	60.0	Libano (1960)	508	13.0	26.0	61.0	Bulgaria (1962)	530	26.8	40.0	33.2	
Gabón (1970)	497	5.8	23.7	67.5	Uruguay (1968)	618	16.5	36.5	48.0	España (1965)	750	17.6	36.7	45.7	**
Jamaica (1958)	510	8.2	30.3	61.5	Chile (1968)	744	13.0	30.2	56.8						
Costa Rica (1971)	521	11.5	30.0	58.5											
México (1969)	645	10.5	25.5	64.0											
Sudáfrica (1965)	669	6.2	35.8	58.0											
Panamá (1969)	692	9.4	31.2	59.4											
Venezuela (1970)	1.004	7.9	27.1	65.0	Argentina (1970)	1.079	16.5	36.1	47.4	Polonia (1964)	850	23.4	40.6	36.0	
Finlandia (1962)	1.599	11.1	39.6	49.3	Puerto Rico (1968)	1.100	13.7	35.7	50.6	Japón (1963)	950	20.7	39.3	40.0	
Francia (1962)	1.913	9.5	36.8	53.7	Países Bajos (1967)	1.990	13.6	37.9	48.5	Reino Unido (1968)	2.015	18.8	42.2	39.0	
					Noruega (1968)	2.010	16.6	42.9	40.5	Hungría (1969)	1.140	24.0	42.5	33.5	***
					Alemania, Rep. Federal (1964)	2.144	15.4	31.7	52.9	Checoslovaquia (1969)	1.150	27.6	41.4	31.0	
					Dinamarca (1968)	2.563	13.6	38.8	47.6	Australia (1968)	2.509	20.0	41.2	38.8	
					Nueva Zelandia (1969)	2.859	15.5	42.5	42.0	Canadá (1965)	2.920	20.0	39.8	40.2	
					Suecia (1963)	2.949	14.0	42.0	44.0	Estados Unidos (1970)	4.850	19.7	41.5	38.8	

Nota: Las participaciones de ingresos de cada grupo percentil se han leído en una curva de Lorenz trazada a pulso y ajustada a los puntos observados en la distribución acumulada. Las distribuciones se refieren a ingresos antes de impuestos. Las cifras del PNB per cápita proceden de archivos de datos del Banco Mundial y se refieren a PNB al costo de los factores en el año indicado y en dólares constantes de los Estados Unidos en 1971.

Los datos utilizados en las tablas de este artículo se han tomado principalmente de Jain, S. y Tiemann, A. 1973. *Size Distribution of Income A Compilation of Data*. Development Research Center Discussion Paper No. 4. World Bank. Washington, D.C.

Nota: Tomada de (1)

* Ingreso hasta US $ 300.

** Ingreso hasta US $ 300 - $ 750

*** Ingreso superior a US $ 750

5.6 Sobre cómo no validar vía "concordancia"

Los ensayos de validación empírica de clasificaciones o tipificaciones han sido hasta la fecha muy escasos. Ello constituye la razón principal para determinar el énfasis general de estos Comentarios.

Entre las pocas pruebas estadísticas explícitas de la bondad de clasificaciones de empresas agropecuarias logradas por "mêtodos tradicionales" (no estadísticos), se encuentra un trabajo de Boussard y Petit (4)*. Los comentarios que siguen se refieren a este trabajo. A pesar de existir críticas, tanto a las bases teóricas, como al proceso de construcción práctica de las clases en el establecimiento de la tipología de las explotaciones, la atención aquí se dirigirá sólo al contenido de la Sección II (Test Estadístico de la Tipología), sección que se agrega como Anexo 1 a este documento.

El primer mensaje-advertencia que sale de la consideración de la metodología de validación empleada por Boussard y Petit, es que la construcción de la/s variable/s "contrastadora/s" o "testigo/s", debe prescindir de la elección de criterios que POR NECESIDAD Y/O DIRECTAMENTE representan diferentes versiones de los propios criterios de tipificación. De lo contrario uno se aproxima en mayor o menor medida a dócimas que no son dócimas realmente, sino meros artefactos de reconstrucción de las clases construídas al inicio del proceso. El resultado natural a esperar es una validación positiva, desde luego.

De hecho esto es lo que sucede en este trabajo. Sin entrar en detalles, basta advertir que el FENOMENO constrastador (representado por medio de cuatro variables alternativas)** es el **"grado de intensidad del sistema"** (intensidad, intensidad de producción, intensidad de las culturas, intensidad de las explotaciones de la región), mientras que el fenómeno productor de la tipología a ser contrastada es... el **grado de intensidad del sistema!** Porque, qué otra cosa están representando sus tres criterios univariados: densidad de la mano de obra, porciento de la SAU irrigable y porciento de viñas y, consecuentemente, sus combinaciones de a dos y sus combinaciones de a tres? Más grave aún: **ambos** conjuntos de criterios, los productores de la tipología y los "contrastadores", se refieren a un mismo tipo especial de intensidad: intensidad en el uso del factor fijo **tierra.** O **sea todo** viene medido en **"algo por unidad de tierra"**: mano de obra por hectárea, tierra irrigable por hectárea, y viñas por hectárea, en lo que hace a las variables representativas del fenómeno productor de la tipología; y cereales y forrajes por hectárea, legumbres por hectárea, legumbres y viñas por hectárea y producto bruto por hectárea, en lo que hace a las variables representativas del fenómeno contrastador. De hecho, entonces, no sólo ambos fenómenos representan esencialmente la misma cosa, sino que en los **denominadores** de todas las variables-cocientes que los representan se encuentra la MISMA variable original, cantidad de tierra, o más precisamente superficie agrícola utilizada (SAU). Lo que obviamente agrava la objeción.

Una forma directa de ilustrar más aún la objeción anterior se logra acudiendo a la misma evidencia empírica suministrada por los autores. Se observa en el Cuadro de página 112 del Anexo que, para el criterio "contrastador" que incluye viñas (tercera columna del cuadro) los valores calculados de los estadísticos Chi-Cuadrado son para los criterios clasificadores que INCLUYEN viñas entre dos y ocho veces los valores calculados para los criterios que EXCLUYEN viñas!

Si uno divide el cuadro de la página 112 del Anexo en tres partes: por un lado las tres primeras líneas correspondientes a los resultados empleando una sola variable clasificadora; una segunda con las

---

* Una traducción del francés de este trabajo se incluye como Anexo a estos Comentarios, por lo que se omitirá aquí una descripción del mismo. En ocasión del Seminario sobre Tipificación de 1975 se tuvo acceso al mismo por gentileza de su Coordinador, Hugo E. Cohan.
Algunas pruebas estadísticas de resultados derivados de la aplicación de métodos estadísticos de tipificación pueden encontrarse en (8), en sección IV.

** Vid. (4), pag. 111.

tres líneas siguientes correspondientes a los resultados empleando cruzamientos de a dos variables clasificadores; y una tercera con la última línea correspondiente a los resultados empleando cruzamientos de las tres variables clasificadores, se observa lo siguiente:

En la segunda porción del cuadro aparentemente los niveles de significación estadística se encuentran mal establecidos*. En la tercera porción del cuadro aparentemente se encuentran mal establecidos los niveles de significación estadística y los grados de libertad**.

Por otra parte, no es sorprendente que los resultados de la última línea, tercera y quintas columnas exhiban el mismo grado de significatividad estadística, como consecuencia de exhibir los mismos valores para el estadístico calculado: una variable-columna es el complemento a uno de la otra! Y este es el único criterio contrastador para el cual la "validación" ensayada, de hecho "valida"

Aún dejando de lado las objeciones metodológicas arriba ensayadas, surge una aún mayor. En el párrafo anterior se ha visto que dos de los resultados "significativos" se reducen en realidad a uno, que es de hecho el único significativo. Y los autores argumentan que el que aparece como "no significativo", lo sería "a nivel ligeramente más elevado". Esto último no tiene sentido. En reciprocidad, entonces, se podría decir que el único significativo, no lo sería "a un nivel ligeramente menos elevado"! Y vuelva a observarse que el único nivel significativo encontrado en el rechazo de la hipótesis nula de independencia, se encuentra para la variable contrastados que INCLUYE viñas, vis-a-vis una clasificación que INCLUYE viñas!

Por todo lo anterior, uno no puede hacer a menos que considerar que el auto-gratificante juicio que dice "De ahí que encontramos una justificación de la hipótesis fundamental de nuestra clasificación", sea por decir lo menos, una simple y llana exageración. Lo que además se complementa muy bien con lo que sigue , que no es sino una típica "ex-post-facto racionalización" de resultados desfavorables, que sin duda deja mucho que desear, metodológicamente. Se llega incluso al extremo de argumentar que un resultado desfavorable se debe al hecho de que "gran parte de las tierras irrigables. . . no fueron efectivamente irrigadas"***, mientras que por otra parte se argumenta que otro resultado desfavorable se debe al "hecho de que las legumbres son probablemente más dependientes del riego que la viña". Además, si el primer argumento es válido, uno debe entonces preguntarse, para qué se incluyó la superficie irrigable como uno de los tres criterios básicos de la clasificación?

En resumen, parece entonces que, a la luz de las objeciones metodológicas y re-evaluación de los resultados hasta aquí expuestos, la apreciación sumaria de los autores, de que "En resumen, parece entonces que la hipótesis fundamental de la clasificación propuesta se halla justificada",**** no se halla justificada.

El espacio dedicado aquí a este trabajo no se hallaría a su vez justificado, si dichas prácticas pudiesen ser adscriptas sólo al mismo. Como el autor de estos comentarios firmemente cree que ello no es así, sino que ellas son generalizadas, es que quiere hacer sonar este toque de atención. En otras palabras, resaltar lo nocivo y anticientífico de los intentos -por otra parte humanos y entendibles- que se pueden hacer, conscientemente o inarvertidamente, por mostrar que "nuestra tipificación es buena"; que el proceso de validación (aparentemente objetivo e impecable) garantiza el éxito de la tipificación ensayada. De allí que el primer crítico de las tipificaciones que se ensayan debe ser el propio tipificador. Y si intenta "validar"

---

* Cf. los valores calculados con los valores de Chi-Cuadrado de tablas, incluídos como segundo Anexo al presente documento.

** Los grados de libertad de Tablas de Contingencia de doble entrada, se determinan como GL = (r – 1) (c – 1), donde r: número de filas en la tabla y c: número de columnas en la Tabla de Contingencia. (8), pp. 108 - 113.

*** Cf. (4), pag. 113 del Anexo No. 1

**** IBIDEM, pág. 113.

el proceso, él también debe constituirse en el primer crítico del propio proceso de validación. La Estadística puede ayudar en ambas etapas, pero ella no puede hacer milagros cuando se la inserta en un proceso que exhibe una base metodológicamente endeble.

Sólo para ilustrar los riesgos de la tendencia señalada en el párrafo anterior, se puede acudir a uno de los trabajos presentados a esta misma reunión, y ya citado anteriormente*. Si bien la discriminación en él ensayada fue hecha a título ilustrativo solamente, no se emite un juicio basado en un dócima o prueba formal, y el juicio emitido es "suave", la conclusión que cierra la aplicación: "La clasificación de las empresas aparece como buena, en la medida en que todas las observaciones presentan una probabilidad de 1.00 de estar clasificadas correctamente"** es metodológicamente no defendible. Ello es así porque, al no existir grados de libertad en el proceso de discriminación (existen cuatro variables empleadas en el análisis y solamente tres observaciones en cada uno de los dos grupos), dichas probabilidades no pueden ser sino uno ("no tienen más remedio" que ser uno). Esto, que el propio autor reconoce, se enfatiza acá para resaltar los riesgos de un mal uso de técnicas.

Para emitir un juicio conclusivo en esta materia, sobre la base de las probabilidades así calculadas (es decir, aparte de, o en adición, a, las dócimas formales de D^2, F y/o Chi-Cuadrado previstas en la literatura, para la consideración de diferencias estadísticamente significativas entre grupos), puede hacerse referencia aquí al área cubierta en (15).

Lo anterior tiene que ver con la concordancia entre una tipificación y determinado principio o criterio. Existe otro tipo de "concordancia": entre agrupamientos o particiones. Los principios que la rigen, y las técnicas aptas para poner de manifiesto su existencia (o ausencia de ella) son similares. Para ilustrar los puntos 4.2.1. (dócimas o tests de independencia) y 4.2.2. (medidas de dependencia) del trabajo de Pedro Ferreira presentado a esta reunión*** en el contexto de evaluación de la similaridad entre particiones, se incluye en el Cuadro No. 3 un sumario de los resultados correspondientes al ejemplo incluído en el punto 4.6.1. de dicho trabajo.

Cuadro 3

Tabla de Contingencia sobre ejemplo de P. Ferreira

Partición I / Partición II	1	2	3	Total
A	3. 1.67 1.06	1. 1.67 0.27	1. 1.67 .27	5
B	2. 3.33 .53	4. 3.33 .13	4. 3.33 .13	10
Total	**5**	**5**	**5**	**15**

* (2), ANEXO, Análisis Discriminante.

** IBIDEM, pág. 78

*** Vid. (6), pp. 84, 87 y 88.

donde, en la primer fila de cada celdilla figura el número de observaciones o frecuencias, en la segunda figura la frecuencia teórica respectiva, y en la tercera el cociente entre el cuadrado de la diferencia entre frecuencia observada y teórica y (con respecto a) la frecuencia teórica.

La suma de los estadísticos incluídos en las terceras filas de todas las celdillas de la tabla arriba, genera el valor del estadístico Chi-Cuadrado (2.39) que, contrastado con los valores de tablas, a los niveles usuales de significación estadística, resulta en el no rechazo de la hipótesis nula de independencia.

Las medidas de dependencia usuales, asociadas a tablas de contingencia de este tipo* exhiben los siguientes valores:

$$\phi^2 = x^2 / n = 2.39/15 = .16 \; ; \; \sqrt{\phi^2} = \phi = .4$$

$$V = (2.39 / (15 \times 1))^{1/2} = \sqrt{.16} = .4$$

$$C = (2.39 / (2.39 + 15))^{1/2} = (2.39 / 17.39)^{1/2} = \sqrt{.137} = .37$$

5.7 Sobre criterios de estabilidad en procesos de validación

Todo lo de la sección anterior tiene que ver, directa o indirectamente, con los **criterios de concordancia** empleados en procesos de validación de procesos de tipificación, que Pedro Ferreira comenta en su trabajo**

En lo que hace a los **criterios de estabilidad*****, ellos son importantes en relación con el punto señalado en la sección III de arriba. La idea de "estabilidad" puede hacerse extensiva a la evaluación de diferentes técnicas alternativas empleadas para un mismo problema de tipificación. Según ellas conduzcan a similares (diferentes) agrupamientos, es decir que las observaciones no "salten" de un grupo a otro a través del empleo de diferentes técnicas, se puede lograr una evaluación de la robustez del proceso tipificatorio que se esté intentando.

En el empleo de la técnica de análisis discriminante, un criterio de estabilidad puede venir dado por la observación de los "saltos" de observaciones de un grupo a otro/s a medida que se vayan ajustando (modificando) los grupos correspondientes a corridas sucesivas del análisis, conforme a los resultados de las mismas. Si bien intuitivamente puede esperarse un proceso de convergencia hacia un equilibrio (estable) a medida que se avanza sucesivamente en el proceso de modificación de grupos**** la propiedad de estabilidad del proceso tipificatorio podría evaluarse por la rapidez manifiesta en dicho proceso de convergencia (ej.: el número de iteraciones sucesivas demandado para lograrla).

La revisión de un trabajo reciente (14) en esta área sugiere como reacción a él, la siguiente advertencia: la consideración de propiedades de estabilidad de un proceso de tipificación que genere grupos alternativos provenientes de diferentes selecciones de atributos o variables, deberá hacerse sobre la base de conjuntos diferenciados de atributos sí, pero que sean representativos del **mismo** tipo de fenómeno

---

* Cf. (8), Anexo 4, pp. 114 - 117.

** Cf. (6), sección 4.2. Concordancia con determinado principio o criterio, pp. 83, 84, 87 y 88.

*** IBIDEM, pp. 84 y 85.

**** Esta propiedad podría/debería investigarse en forma rigurosa.

orientador o tipificador. Es natural esperar poca o ninguna estabilidad cuando se comparan tipificaciones basadas en conjuntos alternativos de variables que responden a, o se identifiquen con, diferentes fenómenos o problemas. Y ello debe mantenerse aún cuando las esencialmente DIFERENTES variables se llamen por el MISMO nombre (o parecido). En este sentido el trabajo arriba aludido constituye una útil referencia*.

5.8 Sobre evaluación de similaridades entre particiones o grupos

A continuación sigue una ilustración de los resultados obtenidos siguiendo el enfoque alternativo propuesto en (6)** para la evaluación de la similaridad entre particiones o grupos.

En la siguiente lista del total de comparaciones de a pares correspondiente al ejemplo suministrado por Pedro Ferreira***, se han subrayado todas aquellas que constituyen "pares similares" (en el sentido que pertenecen al mismo cluster o grupo) (PS):

1-2, 1-3, 1-4, 1-5, 1-6, 1-7, 1-8, 1-9, 1-10, 1-11, 1-12, 1-13, 1-14, 1-15

2-3, 2-4, 2-5, 2-6, 2-7, 2-8, 2-9, 2-10, 2-11, 2-12, 2-13, 2-14, 2-15

3-4, 3-5, 3-6, 3-7 3-8, 3-9, 3-10, 3-11, 3-12, 3-13, 3-14, 3-15

4-5, 4-6, 4-7, 4-8, 4-9, 4-10, 4-11, 4-12, 4-13, 4-14, 4-15

5-6, 5-7, 5-8, 5-9, 5-10 5-11, 5-12, 5-13, 5-14, 5-15

6-7, 6-8, 6-9, 6-10, 6-11, 6-12, 6-13, 6-14, 6-15

7-8, 7-9, 7-10, 7-11, 7-12, 7-13, 7-14, 7-15

8-9, 8-10, 8-11, 8-12, 8-13, 8-14, 8-15

9-10, 9-11, 9-12, 9-13, 9-14, 9-15

10-11, 10-12, 10-13, 10-14, 10-15

11-12, 11-13, 11-14, 11-15

12-13, 12-14, 12-15

13-14, 13-15

14-15

---

* Se agradece a Hugo E. Cohan el envío de este trabajo, para su consideración, en ocasión del Seminario de 1975.

** Cf. (6), p. 87 y 88.

*** IBIDEM, p. 87.

El cociente entre PS y el número total de comparaciones de a pares (N) genera la medida de similaridad entre particiones de Rand (Ra); de manera que se tiene

$$Ra = PS/N = \frac{16}{\frac{15 \times 14}{2}} = 16 / 105 = .15 \quad ;$$

de tal forma que su complemento a uno, empleado por Green y Rao (GYR) es:

GYR = 1 – Ra = 1 – .15 = .85; o, donde PD significa "pares disímiles" (en el sentido de que NO pertenecen al mismo cluster o grupo):

GYR = PD / N = 89 / 105 = .85

Tanto el "Ra" como naturalmente su complemento a uno, el "GYR", exhiben limitaciones como medidas de similaridad. Ello se debe a que sus cocientes se toman sobre el TOTAL de comparaciones de a pares. En búsqueda de una mejor representación de la situación de similaridad entre grupos, puede pensarse que DADO un grupo, se podría comparar el mismo numerador de Ra (o GYR) con el denominador constituído por los POSIBLES pares coincidentes (dado aquel grupo que se toma como punto de referencia). De la misma manera, DADO el OTRO grupo, se podría hacer lo mismo que en el primer caso; y luego tomar como medida representativa de la similaridad entre grupos, un promedio de ambos resultados. Este criterio se hallaría sin duda más cercano al concepto de "probabilidad", ex-ante.

Para el ejemplo que se viene discutiendo, siguiendo el procedimiento aquí propuesto, se tendría:

DADA la partición I : 16/55 = .29
DADA la partición II : 16/30 = .55

de manera que el nuevo indicador, presuntuosamente llamado "MAK", sería:

MAK = (.29 + .53) / 2 = .41

Ex-post, y al menos para este ejemplo particular, dicho indicador genera un valor más cercano a los ya encontrados precedentemente* para las medidas usuales de dependencia asociadas a tablas de contingencia: φ^2, V y C (.40, .40 y .37, respectivamente).

Sin duda convendrá investigar en el futuro el comportamiento y las propiedades de indicadores como Ra y GYR, ya que es de esperarse que estas medidas variarán conforme con el NUMERO de grupos resultantes del proceso respectivo de tipificación. En particular, deberá investigarse su comportamiento en función de dicho número. De igual manera, si pareciera conveniente, será necesario investigar el comportamiento en situaciones de comparaciones MULTIPLES (como opuesto a comparaciones entre simples PARES de particiones o grupos), de indicadores como el aquí propuesto "MAK", o similares que atiendan a la misma idea central.

---

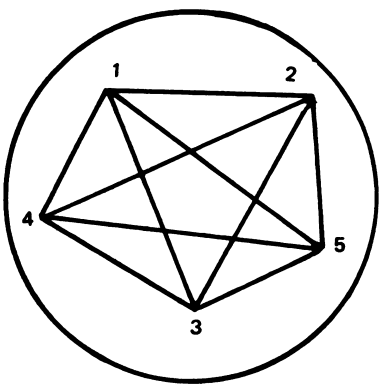
* Cf. precedentemente, en pág. 103.

5.9 Sobre (una ilustración de) grafos aleatorios

La idea de aproximarse al máximo al concepto de probabilidad, y consecuentemente de distribución de probabilidad, que se halla implícita en indicadores del tipo "MAK" propuesto* está explícitamente considerada en la teoría de grafos aleatorios, de la que trata el Apéndice de (6)**. Un intento muy sumario de conectar por medio de una ilustración esta área, con aquellas incorporadas en los puntos 4.2. 4.5. y 4.6.*** del mismo trabajo, consistiría en lo siguiente:

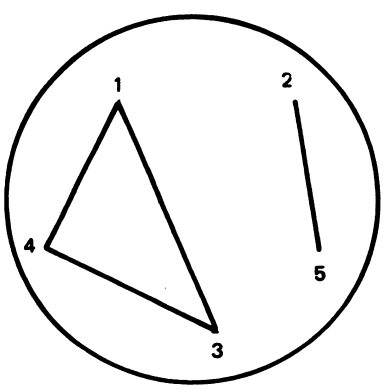
Para el caso del ejemplo en la página 5 de (6), con cinco elementos se tendrían

$C_2^5 = 5! / 2!\,3! = 5 \times 2 =$ 10 posibles "lados", en un grafo como el siguiente:



A un nivel de admisibilidad 2, para el ejemplo aludido, las particiones resultantes se representan en un grafo de cuatro lados, así:

I)



---

* Cf. Sección anterior.

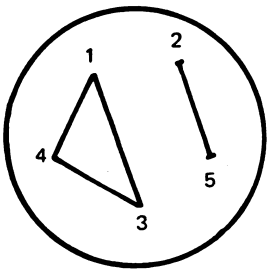
** Cf. (6), Apéndice, pág. 89

*** IBIDEM, pp. 83, 85, 87, respectivamente.

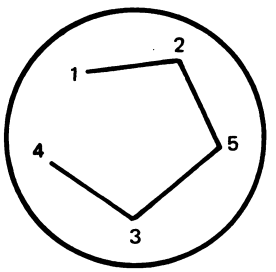
Claro que como el anterior existen, tomando al azar cuatro del total de 10 lados o líneas continuas, un total de

$C_4^{10} = 10 ! / 4 ! 6 ! = 30 \times 7 =$ 210 posibles grafos igualmente probables, así;

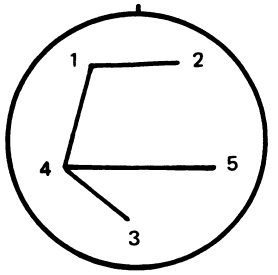
I)



II)



CCX)



De allí que el grafo I que incorpora la partición de análisis de conglomerados jerárquicos a nivel de admisibilidad 2, presenta un elemento de este conjunto de 210 posibles grafos igualmente probables. Y cualquiera de ellos, tomado al azar de entre este conjunto, constituye un "grafo aleatorio".

5.10 Sobre el empleo de correlación canónica en procesos de validación

La marca distintiva de las técnicas "no tradicionales" de tipificación, es que en sus aplicaciones ellas envuelven conjuntos de variables cuyos roles son representar determinados fenómenos de interés, bases para la tipificación que se ensaya. Es decir, se trata de situaciones multivariantes o multivariadas.

Una vez que ello se reconoce, surge naturalmente la recomendación del empleo en procesos de tipificación, de precisamente una técnica multivariante de poco difundido uso en econometría en general, en economía agraria en particular, y más especialmente aún en tipificación de empresas agropecuarias y sus procesos de validación asociados. Se trata de la técnica de Correlación Canónica.

Este no es el lugar apropiado para una descripción del método, de sus propiedades, ventajas y limitaciones. Una descripción sumaria de él y de algunas de sus aplicaciones puede encontrarse en (8) y en mayor detalle en (12) y (13)*.

Ella esencialmente trata de la estimación de relaciones entre dos conjuntos de variables estocásticas, y puede ser empleada con provecho en procesos de validación de tipificaciones bajo diversas circunstancias.

Así, en el estudio de la concordancia de una tipificación con determinado principio o criterio, ella puede ser utilizada aplicándola al estudio de la hipotética relación entre el conjunto de variables originalmente empleada en la tipificación y otro conjunto de variables contrastadoras o testigos. En el caso del trabajo incluído aquí como Anexo, se trataría entonces de la estimación por medio de Correlación Canónica, de la relación entre el conjunto de variables (α, β, γ) y el conjunto de variables que supuestamente representarían el fenómeno contrastador "intensidad de producción del sistema".**

Asimismo esta técnica también puede ser empleada con provecho en combinación con variables binarias o "dummy", cuyo rol sería identificar la hipotetizada existencia de cambios paramétricos, a través de los grupos logrados en un proceso de tipificación, en el tipo de relaciones a que se aludió en el párrafo precedente.

Por último, ella puede también ser empleada con provecho y en forma económica, combinada con la técnica de Análisis Discriminante y, en el caso de carecerse de información original sobre algunos de los conjuntos de variables a ser empleadas, combinada con operaciones intensivas de muestreo.

---

* (8) Apéndice 2, pp. 83-100
(12) pp. 62-65 y 90-95
(13) Diversas secciones

** Vid. la discusión precedente sobre Boussard y Petit.

5.11 Referencias

1 AHLUWALIA, Montek, Sing, **La Desigualdad de Ingresos: Algunos Aspectos del Problema,** reproducido en CIENES (ML/1091 (120), 12.5.76) de "Finanzas y Desarrollo", Volumen 11, Número 3, Septiembre de 1974, pp. 3-9, Tomado del Capítulo 1 de CHENERY, Hollis, AHLUWALIA, Montek S., I, BELL, C. L. G., DULOY, John H., y JOLLY, Richard, **Redistribution with Growth** (Oxford University Press, London, 1974)

2. ALONSO, Alfredo, **Algunas Técnicas de Conglomeración - Su Naturaleza y sus Posibilidades en Tipificación de Empresas,** Reunión Técnica sobre tipificación de empresas agropecuarias, IICA-MAP, Montevideo - 1978, incluido en esta publicación

3. BISIO, Raúl, MARTINEZ, Juan Carlos, y TRIGO, Eduardo, **Problemas Metodológicos y Operativos en la Tipificación de Empresas Agropecuarias: La Experiencia Plan Nacional de Abastecimiento,** Seminario sobre Métodos y Problemas de Tipificación de Empresas Agropecuarias, IICA-MAP, Montevideo, Uruguay, Noviembre 3-7, 1975. IICA, Serie de Informes, Cursos y Reuniones No. 92, Montevideo, Diciembre 1975, Volumen 3, 1, pp. 1-36 y Apéndices No. 1 a No. 3.

4. BOUSSARD, J. M., et PETIT, **Problemes de l'Accession à l'Irrigation,** Institute Nationale de la Recherche Agronomique, 1966. Traducción parcial para fines docentes hecha por Laura Villoria de Kaminsky. La Sección II del Apéndice en dicho trabajo se presenta como Anexo 1 en este Documento.

5. COHAN, Hugo (Ed.), **Seminario sobre Métodos y Problemas en Tipificación de Empresas Agropecuarias,** IICA, Serie de Informes de Conferencias, Cursos y Reuniones No. 92 (Montevideo, diciembre, 1975), Volumen 1, 1.

6. FERREIRA, Pedro, **Algunos Comentarios sobre Evaluación de Clusterings,** Reunión Técnica sobre Tipificación de Empresas Agropecuarias, IICA-MAP, Montevideo, Uruguay, 1978. Incluído en esta publicación.

7. –––––––– **Técnicas Disponibles para Tipificación de Empresas Agropecuarias,** Seminario sobre Métodos y Problemas en Tipificación de Empresas Agropecuarias, IICA-MAP, Montevideo, Uruguay, Noviembre 3-7, 1975, IICA, Serie de Informes de Conferencias, Cursos y Reuniones No. 92, Montevideo, Diciembre 1975, Volumen 1, 5.

8. KAMINSKY, Mario, **Aplicaciones e Ilustraciones de Técnicas Disponibles para Tipificación de Empresas Agropecuarias,** en Cohan, M. (ed) "Seminario sobre Métodos y Problemas en Tipificación de Empresas Agropecuarias", IICA, Serie de Informes de Conferencias, Cursos y Reuniones No. 92, Montevideo, Diciembre 1975.

9. ––––––––, **El Enfoque Probabilístico en Economía y Ciencias Sociales,**. Notas asignatura "Modelos Probabilísticos en Economía y Ciencias Sociales", Curso de Estadística Aplicada al Campo Económico-Social, CIENES 10383 (Santiago de Chile, 1974).

10. ––––––––, **El Problema de la Identificación en Modelos Uniecuacionales y Multiecuacionales,** Notas asignatura "Modelos Probabilísticos en Economía y Ciencias Sociales", Curso de Estadística Aplicada al Campo Económico-Social, CIENES/10833 (Santiago de Chile, 1976).

11. ––––––––, **Comentario** al trabajo "Problemas Metodológicos y Operativos en la Tipificación de Empresas Agropecuarias: La Experiencia Plan Nacional de Abastecimiento", Seminario sobre Métodos y Problemas en Tipificación de Empresas Agropecuarias, IICA-MAP, Montevideo, Uruguay, Noviembre 3-7, 1975. IICA, Serie de Informes de Conferencias, Cursos y Reuniones No. 92, Montevideo, Diciembre 1975, Volumen 3, 1, Comentario, pp. 1 - 4.

12. ––––––––, "Estimación de Hipersuperficies de Producción de Producto Múltiple Libres de Sesgo Empresarial", **Cuadernos de Economía,** Año II, Agosto 1974, No. 33.

13. ––––––––, **The Structure of Production of Multiple Output Dairy in the "Centro Santafesino" Region of Argentina; A Multivariate Analysis,** Tesis Ph.D., Universidad de Wisconsin, 1971.

14. PRETZER, Don D., y FINLEY, Robert M., "Farm Type Classification Systems: Another Look at an Old Problem", **American Journal of Farm Economics,** Vol. 56, No. 1, Febrero 1974.

15. SORUM, M., "Three Probabilities of Misclassification", **Technometrics,** Vol. 14, No. 2, Mayo 1972, pp. 309-316.

5.12 Anexo N° 1

Tomado de Boussard y Petit (Traducción parcial para fines docentes)

Sección II
Test Estadístico de la Tipología

La hipótesis fundamental de nuestra tipología es que el sistema de producción de una explotación depende de los factores fijos de dicha explotación. Nuestras clases deben ser tales que el sistema de producción varíe de modo significativo de una clase a la otra. Y dado que, en las encuestas efectuadas en 1960 y 1961 se cuenta con informaciones sobre el sistema de cultivos para cada una de las explotaciones relevadas, resultaba entonces lógico tratar de verificar la hipótesis fundamental de nuestra clasificación por medio de un test estadístico. La concepción y los resultados de los tests efectuados constituyen el objeto de esta sección.

De acuerdo con la hipótesis a docimar, si una cierta variable x puede ser considerada como característica del sistema de producción, la variabilidad de x debe ser mayor entre las clases que en el interior de una misma clase. Si x se distribuye como normal, un análisis de varianza permitiría efectuar la dócima deseada. En el caso más general, no se puede admitir que x se distribuya como normal; se pueden entonces construir clases sobre x y estudiar la contingencia entre la clasificación así constituída y la clasificación a docimar. Bajo condiciones de supuestos estadísticos poco restrictivos, el estudio de esa contingencia puede ser hecho por medio de un test estadístico de χ^2.

Hecho este análisis, resta ahora elegir x. Cómo entonces caracterizar el sistema de producción? Pensando que el punto esencial en la materia consiste en medir el grado de intensidad del sistema, se puede pensar en la proporción de legumbres y viñas en la explotación, o, por oposición, en la proporción de cereales y forrajes. Otro indicador del nivel de intensidad de la producción de una explotación de la región considerada, es el producto bruto por hectárea (o, para emplear una terminología más rigurosa: el valor medio de la producción final por hectárea de superficie agrícola útil).

Es evidente que ninguna de esas variables es un indicador perfecto del nivel de intensidad de las explotaciones de la región; hemos decidido entonces, efectuar una serie de tests para cada una de estas tres variables; o sea, hemos efectuado la misma serie de dócimas para juzgar en qué medida la clasificación propuesta explica las variaciones de la parte de la superficie de legumbres sobre la superficie total, porque uno de los efectos probables de la irrigación es acrecentar considerablemente las superficies con legumbres.

Para constituir las clases sobre las 4 variables elegidas no existía ninguna consideración económica particular para tomar en cuenta. De todas maneras, el número de clases a retener no debía ser demasiado grande, porque de lo contrario las frecuencias en cada celdilla de la tabla de contingencia serían demasiado bajas. Se sabe en efecto que esta es una condición de validez del test de χ^2. Teniendo en cuenta esas consideraciones, las clases han sido hechas por cuartilos. Sin embargo, cuando el cuartilo caía sobre un valor de la variable común a varias explotaciones, todas esas explotaciones han sido incorporadas en la misma clase lo que hace que la distribución marginal de las variables no sea exactamente uniforme en las cuatro clases.

A fin de analizar la influencia de los diferentes criterios, se calcularon varios χ^2 sucesivamente sobre cada uno de ellos, después sobre el cruce de dos criterios y al final sobre el cruce de los tres criterios. Las tablas de contingencia se incluyen en el anexo 16* y los resultados de los cálculos en el cuadro siguiente:

---

* No reproducido en esta publicación (N. del editor).

Cuadro 1

Valores de los calculados, para diversas clasificaciones

Criterio de clasificación	Variable característica de intensidad de cultivo				Número de grados de libertad
	Producto bruto por hectárea	Proporción de legumbres y viñas	Proporción de legumbres	Proporción de cereales y forrajes	
Densidad de mano de obra (α)	54,62**	15.63*	11.63	12.72*	6
SAU irr./SAU (β)	15,02*	6.40	14.43*	2,89	6
Viñas / SAU (γ)	16,74*	50.02**	9,39	31,47**	6
$\beta\alpha$	47,29*	34,40	45,51*	46,82*	24
$\gamma\alpha$	50,69*	74,28**	45,84*	52,40**	24
$\beta\gamma$	37,08*	65,94**	30,39	45,33*	24
$\alpha\beta\gamma$	92,61	104,98*	98,02*	104,87*	66

NOTA: Los valores marcados con un asterisco (*) indican que el χ^2 es significativo a un nivel de probabilidad del 5%.
Los valores marcados con dos asteriscos (**) indican que el χ^2 es significativo a un nivel de probabilidad del 1%.

El examen de la última línea del cuadro demuestra que se puede rechazar la hipótesis nula de independencia de nuestra clasificación con los tres criterios, con respecto a la distribución de 3 de las 4 variables docimadas a un nivel de significación del 5%. La hipótesis de independencia del producto bruto por hectárea no puede ser rechazada a este nivel, pero lo sería a un nivel ligeramente más elevado. De ahí que encontramos una justificación de la hipótesis fundamental de nuestra clasificación.

Las otras líneas del cuadro permiten analizar con más precisión la influencia respectiva de los diversos criterios de clasificación. La densidad de mano de obra de una "clasificación significativa" (al 5%) para la proporción de cereales y forraje y para la proporción de legumbres y viñas; y una "clasificación muy significativa" (al 1%) para el producto bruto por hectárea.

El test para la proporción de legumbres debe ser interpretado con prudencia, ya que parece que se deslizaron ciertos errores en el procesamiento de las fichas de encuesta. En efecto, en principio se relevaba la superficie en viñas y la superficie total en legumbres y en viñas. Más tarde pareció interesante estudiar las variaciones de las superficies en legumbres. Esto último se obtuvo directamente de las tarjetas perforadas, tomando la diferencia entre las 2 superficies que aparecían en las fichas. Así, se nota que la interpretación de las fichas de la encuesta no era siempre muy fácil en lo que concierne a la superficie en viñas. En particular, parece que la superficie en viñas jóvenes, aún no en producción, no había sido tratada en forma uniforme. Aunque se había detectado esa fuente posible de error, la superficie de viñas ha sido cuidadosamente retomada en cada ficha de cultivo, pero se ha omitido de tomar también cuidadosamente, las superficies de legumbres más las de viñas; de manera que nos contentamos con corregir de acuerdo a la corrección de la superficie en viñas. Además no disponíamos de las fichas de la encuesta, las cuales fueron devueltas a la S.C.P. La satisfacción intelectual que habría procurado la corrección de esos errores, pocos números, que existían, no parecía justificar los costos en tiempo y en dinero que hubiese demandado la recuperación de las fichas, la corrección de tarjetas perforadas y la reconstrucción de los cálculos.

La proporción irrigable de la superficie total no da una "clasificación significativa" para el producto bruto por hectárea y para la proporción de la superficie en legumbres. Acabamos de ver que esto está sujeto a cuidado. De todas maneras es posible que el resultado sea válido porque las legumbres de secano, ajos y algunos melones, son poco importantes. Es normal que se obtenga una clasificación muy significativa para la superficie en legumbres más la de viñas, dado que el criterio de clasificación es la superficie en viñas

Ese resultado no nos enseña gran cosa. Es igualmente normal que esa clasificación sea altamente significativa para la proporción de la superficie de cereales y forrajes, porque cereales, forrajes y viñas ocupan una parte importante del territorio agrícola. Se ha visto más arriba porqué hemos incluído ese criterio en nuestra clasificación, en vista del carácter particular de la viña. Es evidente que los resultados del test no contradicen nuestra hipótesis.

Las clasificaciones obtenidas por el cruzamiento de criterios de a dos, son casi todas significativas o altamente significativas; lo que no sorprende teniendo en cuenta los resultados precedentes. Las únicas que no son significativas comprenden β (la proporción irrigable de ls superficie total).

Ello es bastante sorprendente, dado que a priori parece que las explotaciones que pueden regar una gran parte de su superficie deberían poder obtener rendimientos superiores y entonces obtener un producto bruto más elevado. De hecho, esa anomalía se explica si se considera que una gran parte de las tierras irrigables no fueron efectivamente irrigadas en 1961-62 ni lo son actualmente. Ello fue meramente la manifestación de un fenómeno bien conocido, y general, en las zonas recientemente equipadas para la irrigación (ese fue el caso del valle de l'Arc en 1960-61) a saber, la lentitud con que los agricultores adoptan ese nuevo factor de producción y las técnicas a él asociadas.

No debemos, sin embargo, arribar a la conclusión de que el criterio β es inútil para nuestra clasificación, porque uno de los objetivos de nuestro estudio era precisamente explicar ese rechazo de los agricultores enfrentados a dicha innovación.

Si se comparan la segunda y tercera columnas del cuadro, constatamos que las clasificaciones que incorporan β y no incorpotan γ son mejores para la proporción de la superficie en legumbres, que para la proporción de legumbres más la de viñas. La diferencia de los χ^2 calculados probablemente no es significativamente diferente de cero; pero, mientras tanto,eso se puede explicar por el hecho de que las legumbres son probablemente más dependientes del riego que la viña.

En resumen, parece entonces que la hipótesis fundamental de la clasificación propuesta se halla justificada; dicho de otra manera, el sistema de producción de las explotaciones encuestadas en 1960 y 1961 depende de los criterios utilizados para establecer la clasificación. Por último, la densidad de mano de obra familiar masculina por hectárea de SAU se muestra como el criterio más interesante.

Estos resultados aclaran y a la vez justifican aquellos que han sido obtenidos por otro lado, por medio del modelo regional que vamos a estudiar seguidamente.

5.13 Anexo N° 2

Distribución de Chi-Cuadrado

	p = .750	.900	.950	.975	.990	.995	.999
k = 1	1.323	2.706	3.841	5.024	6.635	7.879	10.83
2	2.773	4.605	5.991	7.378	9.210	10.60	13.82
3	4.108	6.251	7.815	9.348	11.34	12.84	16.27
4	5.385	7.779	9.488	11.14	13.28	14.86	18.47
5	6.626	9.236	11.07	12.83	15.09	16.75	20.51
6	7.841	10.64	12.59	14.45	16.81	18.55	22.46
7	9.037	12.02	14.07	16.01	18.48	20.28	24.32
8	10.22	13.36	15.51	17.53	20.09	21.96	26.13
9	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	13.70	17.28	19.68	21.92	24.73	26.76	31.26
12	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	23.83	28.41	31.41	34.17	37.57	40.00	45.32
21	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	28.24	33.20	36.42	39.37	42.98	45.56	51.18
25	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	45.62	51.81	55.76	59.34	63.69	66.77	73.40
50	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	77.58	85.53	90.53	95.02	100.4	104.2	112.3
80	88.13	96.58	101.9	106.6	112.3	116.3	124.8
90	98.65	107.6	113.1	118.1	124.1	128.3	137.2
100	109.1	118.5	124.3	129.6	135.8	140.2	149.4
x_p	.675	1.282	1.645	1.960	2.326	2.576	3.090

For $k > 100$ use the approximation $w_p = (1/2) \times (_p + \sqrt{2k - 1})^2$, or the more accurate $w_p = k \left(1 - \frac{2}{9k} + x_p \sqrt{\frac{2}{9k}}\right)^3$, where x_p is the value from the standardized normal distribution shown in the bottom of the table.

SOURCE. Abridged from Table 8, p. 131, Pearson and Hartley (1962)

* The entries in this table are quantiles w_p of a chi-square random variable W with k degrees of freedom, selected so $P(W \leq w_p) = p$ and $P(W > w_p) = 1 - p$.

CAPITULO 6

Aportes para la tipificación de establecimientos ganaderos en la zona de Areniscas.

Aportes para la tipificación de establecimientos ganaderos en la zona de Areniscas.

6

Martín Dabezies – CIAAB

Oscar Sarroca – CIAAB

6.1 Introducción

El presente trabajo retoma la información y principales conclusiones de un estudio de establecimientos ganaderos, existentes en la zona de Areniscas de Tacuarembó (1).

A partir del mismo se intenta en el presente documento:

a) verificar si dichos establecimientos clasificados por su tamaño (superficie total), implican también diferentes niveles tecnológicos para cada uno de los grupos conformados.

b) recurriendo a la información relevada en dicho estudio, analizar la existencia de ciertos factores latentes en ella, con vistas a un posible replanteo clasificativo.

Para el cumplimiento de estos fines, se pueden usar algunas técnicas estadísticas o cuasi estadísticas, potencialmente útiles para tipificación. Las mismas tienden a resolver el problema de falta de objetividad de los procesos clasificatorios tradicionales*. Fue entonces en este sentido, que se recurrió al Análisis Discriminante y al Análisis Factorial en Componentes Principales, para abordar los puntos a y b respectivamente, ya mencionados.

El objetivo de presentar en este trabajo los resultados obtenidos, así como también los problemas metodológicos que se presentaron y las conclusiones a que se arribó, es el de aportar con ejemplos prácticos al proceso de difusión y aplicación de estas técnicas que se iniciara con el Seminario realizado sobre el tema, en el año 1975 (3).

Las conclusiones intentan recoger también los aportes hechos por Pedro Ferreira en sus comentarios sobre este documento.

Cabe agregar finalmente el agradecimiento de los autores a Guillermo Artigue por la orientación y el continuo asesoramiento brindado en el uso de las técnicas estadísticas empleadas.

---

* Véase en este estudio el punto 6.2. Qué es tipificar empresas y para qué tipificarlas, del Trabajo de Hugo E. Cohan en (2).

6.2 Antecedentes

A los efectos de precisar el marco de referencia de este trabajo, se describirá brevemente el estudio que le sirvió de base (1).

Entre los objetivos del mismo se destaca el de diagnosticar las diferentes tecnologías que en la zona de Areniscas son aplicadas a la ganadería para carne.

A tales efectos, se realizó una encuesta por muestreo, que relevó 64 establecimientos, todos mayores de 100 hás. Fueron eliminados también, todos aquellos predios que no tenían por actividad principal la ganadería para carne.

Los resultados de esta encuesta fueron analizados por estrato de tamaño, a los efectos de permitir visualizar las posibles variaciones de la tecnología utilizada de acuerdo a la superficie. Es decir, que la hipótesis de vinculación entre superficie y nivel tecnológico está contenida en el trabajo original, aunque en él el análisis es somero.

Se identificaron además, para la zona, ocho "sistemas integrales de producción" (1), definidos fundamentalmente por las distintas combinaciones de actividades dentro de la ganadería.

6.3 Pruebas de la calidad de una determinada clasificación

En el estudio mencionado (1), los establecimientos encuestados fueron agrupados en seis estratos de tamaños.

100	a	199	hás.
200	a	499	"
500	a	999	"
1.000	a	2.499	"
2.500	a	4.999	"

Más de 5.000 hás.

Para cada uno de ellos se analizaron los distintos valores que tomaban las variables elegidas. La mayoría de ellas estaban fundamentalmente orientadas, como ya se dijo, a estudiar los niveles tecnológicos existentes.

En base a ésto se intentó determinar si una clasificación realizada por cortes transversales de la variable tamaño (superficie total) se asociaba o no al uso de distintas tecnologías. En otras palabras: se intentó docimar la hipótesis que agrupar predios por tamaño (superficie) es realmente útil para un estudio de tecnología.

Para ello se recurrió al Análisis Discriminante.

Las variables llamadas "tecnológicas", intentan dejar de manifiesto el nivel tecnológico aplicado a las distintas actividades, dentro de la orientación ganadería. Dado entonces, que los 64 establecimientos encuestados poseen distintas combinaciones de actividades (cría, ciclo completo, lanares, etc.), esas variables no son aplicables por igual a cada uno de ellos.

Para salvar este inconveniente y a los efectos del análisis, el conjunto de establecimientos encuestados se partió en dos sub-conjuntos **para ser estudiados separadamente.** De esta forma se obtuvo un grupo de empresas denominadas "criadoras", lo cual no impide que, además tuvieran otras actividades y otro grupo constituído por empresas "ovejeras" (con majada de cría). Estas últimas también independientemente de tener otras actividades.

De esta forma las variables utilizadas para cada grupo se aplican por igual a cada una de las observaciones.

6.3.1. PRIMERA PRUEBA (Establecimientos Criadores)

a) **Número de grupos y número de variables consideradas.**

Los 6 estratos de tamaño del trabajo original, fueron reagrupados en tres:

Grupo 1: 100 a 499 hás.
Grupo 2: 500 a 2.499 hás.
Grupo 3: Más de 2.500 hás.

De esta forma el número de observaciones quedó distribuído de la siguiente manera:

	Nº de observaciones
Grupo 1	20
Grupo 2	28
Grupo 3	8
	56

Las ocho variables que se utilizaron como indicadoras del nivel tecnológico, fueron las siguientes.

- Porcentaje de mejoramientos
- Dotación total en invierno (UA/há)
- Edad al primer entore (meses)
- Duración del entore (días)
- Diagnóstico de preñez (0 = no, 1 = si)
- Edad de destete (meses)
- Suplementación mineral (0 = no, 1 = si)
- Uso de lombricida (0 = no, 1 = si)

b) **Resultados**

El D^2 de Mahalanobis calculado por el programa discriminante, resultó ser:

$$D^2 = 37,1821$$

El mismo resulta mayor al χ^2 de tablas para los dos niveles de confianza utilizados:

$$\chi^2_{16\ 0.05} = 26.29$$

$$\chi^2_{16\ 0.01} = 32.00$$

Cuadro 1

Evaluación de las funciones de clasificación para cada observación

GROUP 1

OBSERVATION	PROBABILITY ASSOCIATED WITH LARGEST DISCRIMINANT FUNCTION	LARGEST FUNCTION No.
1	0 48	1
2	0.98	1
3	0.48	1
4	0.64	1
5	0.53	2
6	0.94	1
7	0.70	1
8	0.51	1
9	0.74	2
10	0 65	1
11	0.95	1
12	0.59	1
13	0.72	1
14	0.44	1
15	0.45	1
16	0.72	1
17	0.56	3
18	0.56	2
19	0 41	3
20	0.51	1

GROUP 2

OBSERVATION	PROBABILITY ASSOCIATED WITH LARGEST DISCRIMINANT FUNCTION	LARGEST FUNCTION No.
1	0 49	2
2	0.52	2
3	0.49	1
4	0.60	1
5	0.53	1
6	0.48	2
7	0.89	3
8	0.84	2
9	0.86	2
10	0.74	2
11	0.74	2
12	0.81	3
13	0.77	2
14	0.74	2
15	0.49	1
16	0.83	2
17	0.80	2
18	0.54	2
19	0.49	3
20	0.40	1
21	0.52	2
22	0.57	3
23	0.84	2
24	0.83	2
25	0.81	2
26	0.58	1
27	0.44	1
28	0.70	1

GROUP 3

OBSERVATION	PROBABILITY ASSOCIATED WITH LARGEST DISCRIMINANT FUNCTION	LARGEST FUNCTION No.
1	0.67	2
2	0.49	2
3	0.93	3
4	0.68	3
5	0.46	3
6	0.93	3
7	0.98	3
8	0.55	3

Nota: Tomado de la salida de computadora y redondeado a dos decimales.

Por lo tanto, no se puede rechazar la hipótesis de trabajo de que los tres estratos de tamaño en que se agruparon los establecimientos se diferencian entre sí en términos de los indicadores tecnológicos utilizados*

En el Cuadro 1, se presenta la parte final del output del Análisis Discriminante. En el mismo se puede apreciar la conformación de cada uno de los tres grupos que se testearon y la probabilidad asociada a la función que mejor discrimina la observación.

De acuerdo a ésto se comprueba que determinados predios deberían ser reasignados a otro grupo, para mejor estar discriminados por tecnología. Expresándolo en porcentaje se obtienen resultados claramente inquietantes:

	% de observaciones mal clasificadas
Grupo 1	25
Grupo 2	43
Grupo 3	25
Sobre Total	34

c) **Conclusiones**

La primera conclusión se refiere al sentido de la aceptación de la hipótesis de trabajo. El hecho de que el D^2 resultante sea mayor al χ^2 de tablas, significa que los tres grupos conformados a priori, no presentan igualdad de medias. Pero esto no implica que los grupos, considerados separadamente, sean homogéneos en lo que a tecnología se refiere. En otras palabras, no asegura que la clasificación realizada sea "la mejor", con respecto a la homogeneidad de los grupos. Al contrario, la escasa diferencia entre el valor computado y el de tablas, puede indicar una considerable importancia a la varianza dentro de grupos, con respecto a la varianza entre grupos.

Esto se confirma en cuanto a que el programa, por un lado detecta una cantidad considerable de casos mal asignados, y por otro asigna bajos valores a las probabilidades asociadas a las funciones discriminantes.

Una interrogante importante que se planteó, es el verdadero poder diferenciador de las variables elegidas. Para aclarar este punto se intentó aplicar la técnica de Kruskall y Wallis.

La misma no pudo ser desarrollada, debido al alto número de repeticiones en los valores que tomaron las variables elegidas (empates), lo cual le resta poder a dicha prueba.

Siguiendo en el sentido de la exploración de las variables elegidas, se pasó a analizar los valores promedios para cada una de ellas, dentro de cada grupo. Esta información también es calculada por el programa discriminante y se presenta a continuación:

---

* No obstante, como se indica en el trabajo de Alfredo Alonso presentado a esta Reunión, las muestras pequeñas con las que estamos trabajando hacen que (dado un número relativamente alto de variables y grupos) sean dudosos los resultados deducibles de un uso mecánico de tablas χ^2, porque la asimilación de D^2 a χ^2 es una propiedad asintótica.

Valores medios de variables, por grupo

	% Mejor.	Dotac.	Edad Entore	Durac. Entore	Diag- nóstico	Edad Destete	Suplem. Mineral	Lombr.
Grupo 1	2.81	0.67	34.20	171.09	0.0	10.55	0.70	0.20
Grupo 2	2.93	0.53	34.28	129.53	0.10	10.35	0.67	0.35
Grupo 3	5.47	0.69	36.00	125.62	0.37	9.62	0.75	0.75

De acuerdo a ésto se observa que las únicas variables en las cuales, desde el punto de vista biológico, hay diferencias más o menos importantes entre grupos son: % de mejoramientos, diagnóstico de preñez y uso de lombricida.

Como resultado de todo este análisis y en especial reforzado por esta última observación, se puede concluir que si bien las diferencias entre grupos son estadísticamente significativas*, la clasificación no es del todo satisfactoria a la luz de los objetivos del trabajo: diferenciación de "grupos tecnológicos".

En el sentido de seguir explorando respecto a la concordancia de la clasificación con variables de tecnología, se ha sugerido el uso de otras técnicas**. Por ejemplo Correlación Canónica, la cual permite detectar aquellas variables de tecnología (combinaciones lineales de las originales) que se hallan más correlacionadas con variables de tamaño (combinaciones lineales de dichas variables).

6.3.2. SEGUNDA PRUEBA (Establecimientos ovejeros)

a) **Nº de grupos y Nº de variables consideradas.** De la misma forma que en la primera prueba, se definieron los mismos estratos de tamaño. Quedando constituídos de la siguiente manera:

	Nº de observaciones
Grupo 1	8
Grupo 2	15
Grupo 3	5
Total	28

En este caso se pudo utilizar cinco variables como indicadoras de tecnología (contra ocho en el caso de criadores):

- Porcentaje de mejoramientos
- Dotación total en invierno (UA/há)
- Borregas 2d. encarneradas en el total de ovejas de primera encarnerada (%)
- Epoca de encarnerada (mes $\bar{X}$)
- Uso de lombricida (0 = no, 1 = si)

---

* Con el recaudo de la nota al pie en página precedente.

** Las otras técnicas serían las citadas en las secciones 4.2., 4.6. y 4.7., en el trabajo que sobre Evaluación de Clustering, presenta Pedro Ferreira.

b) **Resultados y Conclusiones.** En este caso el Discriminante dio resultado negativo para los objetivos perseguidos. En efecto, el D^2 de Mahalanobis dio menor que el χ^2 de tablas al 0,05% de confianza. Los valores fueron los siguientes:

$$D^2 = 13.56$$
$$\chi^2 = 10.0.05$$

Estudiando el Cuadro 2, se observa que los porcentajes de predios mal clasificados fueron los siguientes:

	% de observaciones mal clasificadas
Grupo 1	50
Grupo 2	46
Grupo 3	20
Sobre Total	43

Estos resultados implican exclusivamente que entre los grupos que se estructuraron, no existen diferencias en términos de las variables tecnológicas consideradas.

6.4 Análisis exploratorio de la información original

6.4.1. Objetivo

Testeado, conforme se explicó en el capítulo precedente, un determinado agrupamiento que se realizó en el Estudio de Areniscas (1) también se replanteó en este trabajo el problema clasificatorio.

Al igual que en el estudio original, interesa analizar para el universo de estudio, qué papel juega la Tecnología en relación a otros "factores", como pueden ser: combinación de actividades dentro de la ganadería, dimensión de la empresa, etc. Para lo mismo se recurrió a la técnica de Componentes Principales (o Análisis Factorial), como técnica exploratoria de la información existente.

Estudiando el conjunto de datos, se postuló que en el mismo podrán identificarse algunos sub-conjuntos de variables con mensajes claros o fácilmente interpretables. Estos "mensajes" podrán resumirse en: Tecnología, Tamaño, Orientación dentro de la Ganadería y Tenencia.

Estos subconjuntos de variables, por hipótesis, contribuirían de manera importante a la diferenciación del total de observaciones.

6.4.2. Tratamiento de las observaciones y variables

Se planteó nuevamente el mismo problema que se tuvo al abordar el Análisis Discriminante. No todas las variables que se poseían originalmente eran aplicables por igual a todas las observaciones. Se entendió que ésto podría introducir ciertas distorciones que dificultaran la interpretación de los resultados.

Por lo tanto ya que la actividad "cría de vacunos" era la más importante para la zona y que, además, la mayoría de las variables están referidas a esta actividad; se tomó como universo de estudio a todos los establecimientos criadores. De esta forma la muestra se redujo de 64 observaciones a 56.

Por último, fueron retenidas para el análisis sólo aquellas variables que presentaban variaciones claras entre las observaciones.

Cuadro 2

Evaluación de las funciones de clasificación para cada observación*

GROUP 1

OBSERVATION	PROBABILITY ASSOCIATED WITH LARGEST DISCRIMINANT FUNCTION	LARGEST FUNCTION No.
1	0.63	1
2	0.48	2
3	0.45	1
4	0.70	1
5	0.63	1
6	0.58	2
7	0.75	3
8	0.47	2

GROUP 2

OBSERVATION	PROBABILITY ASSOCIATED WITH LARGEST DISCRIMINANT FUNCTION	LARGEST FUNCTION No.
1	0.47	2
2	0.60	1
3	0.88	2
4	0.51	2
5	0.63	1
6	0.79	2
7	0.77	3
8	0.83	2
9	0.73	2
10	0.45	1
11	0.39	1
12	0.66	2
13	0.52	2
14	0.76	3
15	0.68	1

GROUP 3

OBSERVATION	PROBABILITY ASSOCIATED WITH LARGEST DISCRIMINANT FUNCTION	LARGEST FUNCTION No.
1	0.58	2
2	0.79	3
3	0.75	3
4	0.76	3
5	0.75	3

* Tomado de la salida de computadora y redondeado a dos decimales.

Quedaron de esta forma, 24 variables que agrupadas de acuerdo a la hipótesis de trabajo, se presentan seguidamente:

Tecnología:

1. % de Mejoramientos
2. Dotación total en invierno (UA/há)
3. Edad de entore (meses)
4. Duración del entore (días)
5. Diagnóstico de preñez (0 = no, 1 = si)
6. Edad de destete (meses)
7. Suplementación mineral (0 = no, 1 = si)
8. Uso de lombricida (0 = no, 1 = si)
9. % de Parición
10. Nº de tractores/hás. x 1.000
11. Uso de crédito (0 = no, 1 = si)

Tamaño:

12. Superficie total (hás)
13. Nº de potreros
14. Superficie promedio por potrero (hás)
15. Mano de obra total (E.H./há x 1.000)
16. Mano de obra familiar (% sobre el total)

Orientación:

17. Ovinos/vacunos
18. Venta de Terneros de destete (0 = no, 1 = si)
19. Recría de sus propios terneros (0 = no, 1 = si)
20. Compra para Recriar (0 = no, 1 = si)
21. Hace inverne (0 = no, 1 = si)

Tenencia:

22. Propietario (0 = no, 1 = si)
23. Propietario - Arrendatario (0 = no, 1 = si)
24. Arrendatario (0 = no, 1 = si)

A la variable tenencia, se le dio el tratamiento que se indica anteriormente transformándola en tres variables binarias (var. 22, 23 y 24).

6.4.3. Resultados

a) **Estudio de las correlaciones entre variables.** El programa de Componentes, calcula en primer término la matriz de correlación entre variables. En el Cuadro 3 se presenta un resumen de la misma, elaborado con límite de confianza igual a 1%.

Del estudio de esta matriz, resultan claros dos hechos:

El primero de ellos, es que las variables consideradas a priori como indicadoras de tamaño presentan entre sí correlaciones significativas, a los dos niveles de confianza usados. Dentro de éstas, la variable Nº de potreros, aparece además correlacionada con variables tecnológicas: diagnóstico, uso de lombricida, duración del entore y edad de destete. Esto se puede explicar por el hecho de que, el Nº de potreros también puede ser tomado como un indicador del nivel de tecnología.

La superficie promedio por potrero se correlaciona, además, positivamente con orientación inverne.

La segunda evidencia clara es que, a diferencia de las anteriores, las variables tecnológicas no aparecen tan claramente correlacionadas entre sí. Las únicas asociaciones positivas importantes se dan entre: uso de crédito y suplementación mineral, diagnóstico y uso de lombricida y dotación total en invierno y % de parición.

Dentro de las variables indicadoras de orientación, la única asociación que aparece es la inversa entre venta de terneros y recría propia.

Hasta aquí llegan las conclusiones de analizar correlaciones parciales, realizado como punto previo a un análisis multivariado más poderoso*.

b) **Estudio de las correlaciones entre factores y variables.** En el Cuadro 4 se presentan 10 de los 24 factores de la matriz, factores-variables, sin rotar, y los porcentajes acumulados de la varianza total que explica cada uno de ellos.

A los efectos de una mejor interpretación de los valores de las correlaciones, se estableció como límite inferior de significación de los mismos, el valor de tablas correspondiente al 1% con 54 grados de libertad (Nº de observaciones, menos 2).

Un primer comentario que debe hacerse, es indicar los bajos valores de la varianza explicada por cada factor.

En segundo lugar, la interpretación de los factores no resulta muy evidente.

Tomando los primeros diez, el único factor que puede definirse claramente, es el primero. Se puede hacer por las variables más fuertemente correlacionadas con él, que son variables de "tamaño". Junto a éstas, aparecen asociadas en el mismo sentido, algunos indicadores de uso de prácticas Tecnológicas más avanzadas.

En el resto de los factores aparecen asociaciones no muy claras entre determinadas orientaciones, algunas variables de tecnología y otras de tamaño.

En el intento de clarificar la significación de estos factores, se hicieron dos nuevas corridas del programa, la primera rotando los primeros 9 factores que explican el 74.5% de la varianza total. La segunda, rotan los 7 primeros, los cuales explican el 66%.

Estos resultados se presentan en los Cuadros 5 y 6 respectivamente. Para facilitar su interpretación, se retienen sólo los valores de las correlaciones significativas. Las mismas, al 1% y con 54 grados de libertad, son todas aquellas que están por encima del valor 0.34.

Con la primera rotación (Cuadro 5) se logra clarificar en parte el significado de algunos factores.

El factor 1, queda más identifcado con "tamaño", ya que aumentan las correlaciones correspondientes a esas variables y desaparecen o disminuyen algunas de "tecnología". Este factor explica el 18.4% de la varianza total.

El factor 2, aparece ahora relacionado con variables indicadoras de un uso de tecnología avanzada (por ej.: diagnóstico de preñez, % de mejoramientos, suplementación mineral, uso de lombricida). Este factor explica por si sólo el 11.1% de la varianza y junto con el primero, el 29.5%.

---

* Adviértase que el análisis multivariado efectuado, se inicia a partir de la matriz de correlaciones simples. Ambos enfoques tienen, por consiguiente, problemas derivados de correlacionar variables cualitativas entre sí y con respecto a cuantitativas. Aún no disponemos de un método operativo que resuelva los problemas metodológicos derivados de las muy frecuentes variables cualitativas, como no sea el expediente (no siempre de fácil aplicación) de cuantitivizarlas.

Cuadro 3

Matríz de correlaciones entre variables

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
% Mejoramientos	1																								
Dotación	2																								
Edad Entore	3																								
Duración Entore	4																								
Diagnóstico	5																								
Edad Destete	6																								
Suplementación	7																								
Lombricida	8																								
% Parición	9																								
Tractor/Há.	10																								
Crédito	11																								
Superficie Total	12																								
No. Potreros	13																								
Superficie $\bar{X}$/Potrero	14																								
Equ. Hombre/Há.	15																								
% M. de O. Familiar	16																								
Ovinos/Vacunos	17																								
Vende Terneros	18																								
Recría propia	19																								
Recría compra	20																								
Inverne	21																								
Propietario	22																								
Propiet. - Arrendat.	23																								
Arrendatario	24																								

% Mejoramientos	1																								
Dotación	2									.38															
Edad Entore	3																−.34								
Duración Entore	4								−.34								.37								
Diagnóstico	5								.46					.38											
Edad Destete	6															.43	.43								
Suplementación	7											.40													
Lombricida	8				−.34	.46								.39											
% Parición	9	.38																							
Tractor/Há.	10																								
Crédito	11							.40																	
Superficie Total	12													.70	.83	−.42	−.60								
No. Potreros	13					.38			.39				.70			−.36	−.63								
Superficie $\bar{X}$/Potrero	14												.83			−.41	−.54					.37			
Equ. Hombre/Há.	15						.43						−.42	−.36	−.41		.63								
% M. de O. Familiar	16				.37		.43						−.60	−.63	−.54	.63									
Ovinos/Vacunos	17			−.34																					
Vende Terneros	18																			−.55					
Recría propia	19																	−.55							
Recría compra	20																								
Inverne	21														.37										
Propietario	22																						−.57	−.74	
Propiet. - Arrendat.	23																						−.57		
Arrendatario	24						−.35																−.74		

* Para ayuda visual, se excluyen correlaciones no significativas al 1%. Los valores están redondeados a dos decimales.

Cuadro 4

Matriz de factores (sin rotar, 24 factores)*

Porcentaje de varianza acumulada para los 1ros. diez factores:

		18,4	29,5	39,0	47,0	54,5	60,1	65,5	70,1	74,5	78,1
		F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}
% Mejoramientos	1		−.39			.44	−.51				
Dotación	2			.61							
Edad Entore	3		.39		−.39						
Duración Entore	4	−.53						−.36			.39
Diagnóstico	5	.55									
Edad Destete	6	−.47			−.54	.35					
Suplementación	7				−.64						
Lombricida	8	.64									
% Parición	9	.42							.61	.37	
Tractor/Há.	10		−.49					−.69			
Crédito	11	.36	−.42								
Superficie Total	12	.76	.38								
No. Potreros	13	.75								−.34	
Superficie X̄/Potrero	14	.59	.46	−.34							
Equ. Hombre/Há.	15	−.63		.37							
% M. de O. Familiar	16	−.80		.34							
Ovinos/Vacunos	17						.67				
Vende Terneros	18				−.37	−.66					
Recría propia	19		.50			.57					
Recría compra	20		.45					−.46			
Inverne	21					.43				.48	.36
Propietario	22		−.50	−.74							
Propiet. - Arrendat.	23		.46	.35					.37		
Arrendatario	24			.60	.41						

* Para ayuda visual, se excluyen correlaciones no significativas al 1%. Los valores están redondeados a dos decimales

Los factores 3 y 4 aparecen fuertemente marcados por variables de tenencia. Hasta este último, se está con un 47% de la varianza explicada.

El factor 5, se vincula claramente a orientación dentro de la ganadería: ciclo completo o cría. La varianza explicada hasta aquí es del 55%.

El factor 6, se asocia a orientación ovina o vacuna. A mayor cantidad de ovinos, mayor dotación total en invierno.

El factor 7, pone de manifiesto la posible relación entre el porciento de mejoramientos y la existencia de tractor.

El factor 8, expresa que la dotación está directamente correlacionada con el porcentaje de parición.

Por último, el factor 9, se puede definir como "orientación invernada", la cual aparece asociada directamente a compra de animales para recriar, mayor tamaño (sup. total) y mayor superficie promedio por potrero.

La segunda rotación, Cuadro 6, con la cual se buscaba que aportara una mayor claridad, no resultó útil a este propósito. En efecto, los dos primeros factores (tamaño y Tecnología) resultaron más claros, pero en cambio, los restantes perdieron identidad.

6.4.4. Conclusiones

Atendiendo a la hipótesis de la cual se partió, al encarar el análisis Factorial, se puede decir que en líneas generales la misma se confirma. En efecto, basándose fundamentalmente en el Cuadro 5, se vió que aparecen más o menos claramente, los factores: tamaño, tecnología, orientación y tenencia.

El factor tamaño aparece como más importante que el tecnológico ya que aparece más claro y además explica un 18.4% de la varianza total, frente a un 11.1% del segundo.

El análisis más detallado de los factores lleva a la conclusión de que la conformación que a priori se pensó de cada uno de los factores, en algunos casos se confirma (tamaño) pero en otros no.

Esto último está ejemplificado por el factor tecnología. En efecto, las variables indicadoras de nivel tecnológico no aparecen agrupadas en un sólo factor, sino que tienden a distribuirse en más de uno. Resulta difícil pensar, entonces, en un factor "tecnología" único compuesto invariablemente por un determinado número de indicadores. Esto resulta así para las observaciones y las variables que se estudiaron.

En este mismo sentido, se concluye que, determinadas asociaciones encontradas entre variables tecnológicas y otros factores (por ej.: orientación), no necesariamente deben de poseer una explicación lógica desde el punto de vista físico. Las mismas pueden resultar del azar, para el universo estudiado.

En cuanto al factor tenencia su interpretación definitiva se hace difícil. La causa de esta circunstancia se cree que está en el tratamiento metodológico que se le dio a la variable. De todas formas se cree importante a los fines de esta reunión técnica presentar el problema tal cual fue analizado hasta este momento.

En cuanto al aporte que el Análisis Factorial por Componentes hace a un replanteo clasificatorio, la más importante conclusión a la que se arriba, es que las características particulares que ya se comentaron respecto al factor tecnología, hacen pensar en una reformulación del papel que juega el mismo en el grupo de empresas estudiado.

La tecnología por sí sola o incluso junto con el factor tamaño, no serían suficientes a los fines de una futura clasificación. Esto se fundamenta en los porcentajes de la varianza total explicada por los factores que las representan.

Cuadro 5

Matriz rotada de factores (9 factores)*

		F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9
VARIABLE	1		−0.60					−0.41		
"	2						0.36		0.65	
"	3						−0.57	0.43		
"	4	−0.40					−0.43	−0.45		
"	5	0.41	−0.51							
"	6	−0.47		−0.38	−0.44					
"	7		−0.80							
"	8	0.43	−0.39							
"	9								0.88	
"	10							−0.84		
"	11		−0.61							
"	12	0.77								0.40
"	13	0.79								
"	14	0.59								0.67
"	15	0.70			−0.34					
"	16	−0.88					0.86			
"	17									
"	18					−0.83				
"	19					0.82				
"	20				−0.37					0.36
"	21					0.34				0.71
"	22			−0.80	0.51					
"	23				−0.93					
"	24			0.93						

* Para ayuda visual, se excluyen correlaciones no significativas al 1%. Los valores están redondeados a dos decimales.

Cuadro 6

Matriz rotada de factores (7 factores)*

		F_1	F_2	F_3	F_4	F_5	F_6	F_7
VARIABLE	1		−0.80					
"	2							0.76
"	3						0.75	
"	4	−0.38						−0.63
"	5		−0.60					
"	6				−0.80			
"	7		−0.54		−0.53			
"	8	0.38	−0.50					
"	9							0.36
"	10		−0.39				−0.67	−0.37
"	11		−0.48					0.37
"	12	0.87						
"	13	0.67	−0.35					
"	14	0.83						
"	15	−0.64			0.39			
"	16	−0.83						
"	17						0.68	0.42
"	18					−0.83		
"	19					0.80		
"	20			0.58				
"	21					0.50		
"	22			−0.89				
"	23			0.61	−0.54			
"	24			0.58	0.60			

* Para ayuda visual, se excluyen correlaciones no significativas al 1%. Los valores están redondeados a dos decimales.

CAPITULO 7

Una tipificación de predios lecheros. Uso de técnicas estadísticas en su prueba y reconsideración.

Una tipificación de predios lecheros. Uso de técnicas estadísticas en su prueba y reconsideración.

7

Heraclio Pérez — SEE - DIEA

7.1 Introducción

En este documento preparado para la Reunión Técnica se presentan una tipificación usual de empresas, hecha dentro del programa regular de trabajo de la Subdirección de Estudios Econométricos (SEE-DIEA) y algunos estudios más elaborados sobre el caso.

La tipificación se refiere al agrupamiento de predios lecheros en base a información generada por una encuesta de la que se retuvieron 52 casos válidos. A esta tipificación se la denomina "usual", en cuanto se hizo en función de una variable (lts. leche/Há.) sobre la cual los técnicos participantes acordaron un juicio en cuanto a:

- su poder discriminatorio en general, y
- el rango de variación a partir del cual la variable puede ejercer ese poder.

Los agrupamientos resultantes permitieron definir cuatro predios tipos, a ser modelizados en proyectos dedicados a analizar la rentabilidad de prácticas a-priori superiores a las predominantes.

En los capítulos 7.2. a 7.4. de este documento de discusión se resumen marco de referencia, objetivo, mecánica informativa y resultados de esa tipificación.

La parte de mayor interés para esta Reunión, no obstante, comienza a partir del capítulo VI, previa definición de variables en el capítulo 7.5.

En los capítulos 7.6. a 7.8., se presentan los resultados de aplicar a la tipificación "usual" una prueba de Análisis Discriminante y, luego, los generados al emplearse Análisis de Correlación Simple y Principales Componentes como técnicas para repensar desde el inicio todo el proceso. La tipificación que se presenta en el capítulo 7.4. fue hecha antes de conocerse nuevas propuestas metodológicas. La discusión de los capítulos 7.6. a 7.8. es, entonces, una aplicación con algún sentido de prueba de hipótesis y, principalmente, con interes en explorar tecnicas que hubieran permitido una tipificación más objetiva.

Se estima que la experiencia reunida puede ser útil para reforzar conceptos acordados por los participantes del Seminario que sobre el tema se realizó en 1975, también conjuntamente entre el IICA y la DIEA.

Se agradece el asesoramiento continuo de Guillermo Artigue en la aplicación de técnicas estadísticas. Comentarios de Mario Kaminsky fueron utilizados para la revisión de este documento. La colaboración de Artigue y Kaminsky no es desinteresada, dado que ambos han hecho repetidos esfuerzos por lograr un buen uso de la estadística en tipificación. Es de lamentar que no pueda comprometerse más su nombre en esta presentación, si bien muchos de sus comentarios fueron empleados en la revisión del trabajo original.

7.2 Objetivo del estudio de predios lecheros

El plan de investigación de la S.E.E. incluye el estudio de las principales zonas agroeconómicas del país con el objetivo de analizar la viabilidad económica de sistemas de producción avanzados en relación a los que predominan actualmente. Estos sistemas "avanzados" llevan a cabo combinaciones técnicas caracterizadas, en general, por un mayor uso de insumos que el modal en la explotación de cada rubro y permiten obtener rendimientos físicos por unidad de superficie mayores que los predominantes. El estudio económico requiere, previamente, el reconocimiento de la viabilidad técnica de estos sistemas, corroborada en base a la experiencia de algunos productores y/o la experimentación científica.

Como primera etapa para esta tarea, debe ser relevada información sobre los recursos disponibles por los predios. En particular: características de los suelos, actividades desarrolladas y niveles técnicos empleados.

Una vez sistematizada esa información, la segunda etapa consiste en la elaboración de modelos representativos de los sistemas de producción predominantes y de sistemas más intensivos.

Dentro de dicho plan de investigación, se encaró el estudio de la Cuenca Lechera de Montevideo, dada su importancia dentro de la producción lechera nacional y del sector agropecuario en su conjunto.

7.3 Fuentes de información

Para el estudio de esta Cuenca, se emplearon las siguientes fuentes de información:

- Programa de Estudio y Levantamiento de Suelos
- Censos agropecuarios
- Encuesta de Conaprole del año 1968
- Plan Piloto de Inseminación Artificial del Centro de Investigaciones Veterinarias
- Registros y encuesta del Servicio de Registros del Plan Agropecuario (SERPA)
- Profesionales que efectúan asistencia técnica en la zona
- Encuestas a productores
- Centro de Investigaciones Agrícolas
- Registros de producción de algunos predios

Por su importancia para esta presentación, en este capítulo se detalla una de estas fuentes de información: Encuesta a los productores.

El objetivo primordial de la encuesta fue el de obtener información en profundidad sobre técnicas en aplicación a nivel comercial.

A estos efectos, se analizó con técnicos vinculados al sector lechero qué información relevar a través de la encuesta. Se tuvieron especialmente en cuenta los siguientes elementos:

- Aspectos técnicos mal conocidos a nivel de explotación comercial.
- Limitaciones de la fuente de información (productores) sabiendo que la mayoría no lleva registros.
- Limitaciones inherentes a la encuesta, ya que a través de una sola entrevista se trataría de recomponer un ciclo productivo (año).
- Dificultad de obtener datos económicos, por falta de registros contables y prejuicio de los entrevistados.

Se confeccionó entonces un formulario extensivo, pese a saberse previamente que en la mayoría de los casos relevados no se iba a obtener toda la información solicitada. Esta información se refirió exclusivamente a aspectos físicos de la explotación (insumos, manejo, producción). La formulación de presupuestos quedó para ser realizada en base a los coeficientes técnicos detectados en la encuesta, a combinarse con precios de fuentes diversas.

El trabajo de campo se llevó a cabo durante los meses de julio y agosto de 1975. La elección de las explotaciones a encuestar se hizo con el asesoramiento de técnicos vinculados a la zona y con los siguientes criterios:

- Cubrir adecuadamente el rango de niveles de producción que se obtienen en la Cuenca.
- No tomar en cuenta predios que realizan prácticas de producción poco difundidas en la Cuenca y que se deben a circunstancias especiales, como ser: ubicación geográfica, combinación de la lechería con rubros que raramente la acompañan, alimentación con subproductos industriales poco comunes.
- Descartar aquellas explotaciones que, por estar realizando circunstancialmente transformaciones técnicas importantes, tienen en el momento un manejo del rodeo muy especial.
- No dar al tamaño (superficie) del predio carácter excluyente, por entenderse que los niveles técnicos pueden ser relativamente independientes de él y que una restricción de este tipo reduciría la eficacia de la encuesta en otros aspectos.*

7.4 Definición de predios tipo

En el presente capítulo se explican la mecánica y los resultados de la tipificación que se hizo sin técnicas estadísticas. Al procesarse los formularios se eliminaron aquellos que tenían los problemas ya citados, así como otros en los que no había información sobre aspectos claves o donde la lechería no era el rubro principal.

---

* Más adelante (Cuadro No. 2) se advertirá que en los datos retenidos como representativos de cada estrato, variables tales como litro/há. y porcentaje de praderas nuevas se vinculan inversamente con indicadores de tamaño (superficie vacas masa). Es decir que puede ser errónea la previsión de independencia entre tecnología y tamaño.

A los efectos de la agrupación para la posterior definición de predios tipo para distintos niveles tecnológicos, se consideró a la producción de leche por hectárea como variable preponderante. Ella representa el principal objetivo, aunque no el único, del productor lechero en cuanto a resultado físico.

Debe reconocerse, sin embargo, que para obtener una misma producción de leche es posible llevar a cabo combinaciones técnicas muy variadas. A través de la encuesta se encontró esta realidad a nivel de predios, obteniéndose producciones iguales en predios con uso del suelo muy distinto, diferente número y tamaño de potreros, sanidad, manejo, etc.

Para definir el número de estratos y el rango correspondiente a cada uno, se tomó en primer lugar un límite inferior de 600 lts/há., entendiéndose que establecimientos con rendimientos más bajos tienen problemas especiales cuyo análisis escapa a las intenciones del estudio.

Si bien no existen series estadísticas de varios años para rendimientos en producción de leche por hectárea, se juzga que en la Cuenca esta producción fluctúa entre 650 y 850 lts., dependiendo de las condiciones climáticas y de precios de cada año. Se estimó que a nivel de predio predominarían rendimientos entre 600 y 900 lts/há. Esto fue corroborado por técnicos con experiencia en la actividad, ya que rendimientos de 1.000 lts/há. o más sólo son obtenidos por los escasos predios que utilizan técnicas más intensivas que las predominantes.

Fijado ese estrato representativo de la situación predominante, quedó por resolver cómo agrupar los predios con tecnología y rendimientos superiores. Con los mismos criterios que en el estrato anterior, se consideró que admitiendo rangos de hasta 300 lts/há. hay cierta uniformidad en las técnicas utilizadas y en el nivel de intensidad de éstas. Diferencias mayores de 300 lts/há. resultarían así de sistemas distintos.

Se definió entonces un estrato de 901 a 1.200 y otro de 1.201 a 1.500 lts/há., agrupándose, dado el bajo número de casos disponibles, los establecimientos con más de 1.500 lts/há. En el cuadro 1 se presenta el resultado final de este agrupamiento.

A fin de caracterizar un predio tipo para cada estrato, se siguió el criterio de utilizar el promedio del grupo en aquellas variables que presentaban uniformidad dentro del mismo, mientras que en aquellas en que la información era deficiente o presentaba mucha variación se tomaron los valores más frecuentes.

Cuadro 1

Información sumaria sobre la tipificación

Estrato (lts/há)	Número de casos	Promedio de lts/há.
600 – 900	18	779
901 – 1.200	18	1.059
1.201 – 1.500	11	1.325
Más de 1.500	5	1.850

Esta fue, en esencia, la tipología retenida*.

Al realizarse en Montevideo el Seminario sobre Métodos y Problemas en la Tipificación de Empresas Agropecuarias quedaron dos mensajes considerados por técnicos de la S.E.F. como importantes para esta tipificación. A saber:

1) Caben serias dudas en cuanto a tipificaciones hechas por corte transversal del universo en función de una variable (producción de leche/há., en nuestro caso), y

2) Existen técnicas con gran potencial para testear una dada tipificación, para mejor explorar el universo de variables en cuestión y para, incluso, tipificar.

Los avances logrados al entrar en este proceso más novedoso son el tema central de esta presentación.

Cuadro 2

Valores de las variables para cada uno de los predios tipo*
Tipificación según rangos de producción de leche por hectárea

	TIPOS:	I	II	III	IV
1	Lts/Há.	779	1.059	1.325	1.850
2.	% Praderas Permanentes	3,0	5,7	15,8	36,0
3.	1er. Entore (meses)	36	32	28	24
4.	Carga	0,95	1,07	1,29	1,52
5.	Lts/Vaca	7,4	9,1	9,3	10,1
6	Hás/Potrero	18,3	15	13	5
7	% Unidad lechera Vacas/Vacas total	66	63	61	72
8	Superficie Total	201	194	152	75
9.	Superficie Lechera	201	194	152	75
10.	Vacas Masa	93	96	86	54
11	% Cultivos Comerciales	0	0	0	0
12	% Praderas Nuevas	1,8	2	5,3	10,4
13	Número de Potreros	11	14	15	16
14.	Cuota Total	322	394	359	228
15.	% Alfalfa	5,9	6,1	10,5	16
16.	Cuota/Hectárea	1,60	2,03	2,36	3,04
17	Ración (grs./litro leche)	250	275	200	160
18.	Vacas Producción/Vacas Secas	1,45	1,82	2,31	2,38
19.	% Cultivos Anuales	13	19	11	18
20.	% Praderas Viejas	3,6	2	0	0

* Valores medios o modales, según se indicó en el texto. Las variables se explican en el Capítulo 7.5

* Mayor detalle sobre los tipos se presenta en el Cuadro No. 2.

7.5 Variables empleadas para el uso de técnicas estadísticas

Las 20 variables que se emplearon para Análisis Discriminante y Principales Componentes se describen a continuación:

1. **Producción por hás.** — Producción de leche total en el año dividido entre la superficie lechera.
2. **Porcentaje de praderas** — Area de praderas permanentes que han sido efectivamente utilizadas en el año y mantienen la composición botánica y productividad similar a la original y destinada al rodeo lechero, dividida por la superficie lechera. No incluye alfalfa pura.
3. **Edad 1er. entore** — Edad promedio (en meses) en que las vaquillonas reciben su primer servicio.
4. **Carga (UL/há)** — Unidades Lecheras totales divididas entre la superficie lechera.
5. **Producción por vaca por día** — Producción promedio anual (litros) por vaca en ordeñe por día.
6. **Superficie promedio por potrero** — Superficie lechera dividida entre el número de potreros.
7. **Porcentaje UL vacas/UL totales** — Unidades Lecheras (requerimientos nutritivos) de las vacas masa divididas entre las Unidades Lecheras Totales del rodeo lechero.
8. **Superficie total** — Area del predio entendido como unidad técnica económica. Cuando un productor (empresa) explota más de una unidad de producción: 1) Se tomó el conjunto si fue imposible separar la utilización de recursos; 2) Se tomaron como independientes cuando se pudieron asignar claramente los recursos y producción de cada una. En todos los casos en que existían animales a pastoreo se estimó el área que ocupan y se la tuvo en cuenta.
9. **Superficie lechera** — Area, dentro del predio o en pastoreo, destinada a los animales del rodeo lechero (vacas lecheras y reemplazos). Es igual o menor que la superficie total.
10. **Vacas masa** — Suma de vacas en producción y secas que integran el rodeo lechero.
11. **Cultivos comerciales/hás. totales** — Area destinada a cultivos anuales con destino a la venta, dividida entre la superficie total.
12. **Porcentaje de praderas y alfalfa nuevas** — Area de praderas permanentes y alfalfa pura sembradas en el último año dividida por la superficie lechera.
13. **Número de potreros** — Cantidad de potreros permanentes destinados al rodeo lechero.
14. **Cuota** — Litros de cuota diarios que el productor tiene asignados en Conaprole.
15. **Porcentaje de alfalfa** — Area de cultivos de alfalfa pura efectivamente utilizada dividida por la superficie lechera.
16. **Cuota por hás.** — Cuota dividida por la superficie lechera.
17. **Ración, gramos/lt.** — Consumo total anual de concentrados, por las vacas en producción (en gramos) dividido entre la producción total anual de leche (en litros).
18. **Vacas Producción/Vacas Secas** — Número de vacas en producción respecto al número de vacas secas del rodeo en el momento de la encuesta.
19. **Porcentaje de cultivos anuales (C.A.)** — Area de cultivos forrajeros anuales destinados a la producción lechera dividido por la superficie lechera.
20. **Porcentaje de praderas y alfalfa viejas** — Area de praderas permanentes y alfalfa pura que se encuentran degradadas (han perdido la composición botánica original).

7.6 Prueba de la clasificación mediante análisis discriminante

Se aplicó Análisis Discriminante en la clasificación que se había realizado por métodos no estadísticos con los predios lecheros encuestados. A este fin debieron ser eliminados tres de ellos, sobre los que no se disponía de información para las 20 variables que se consideraron en esta etapa, reduciéndose los grupos a 18, 16, 10 y 5 tambos cada uno.

El reducido número de casos (49) probablemente requeriría retener un subconjunto de 5 ó 6 variables, en vez de las veinte que se emplearon, a efectos de dar mayor posibilidad de manifestarse a las diferencias entre grupos. Esto es un comentario a posteriori, motivado por una mayor experiencia y comentarios críticos en el uso de Análisis Discriminante.

Como resultado del análisis que se realizó, surgió que todos los predios estaban clasificados en el grupo más adecuado, bajo los criterios de estratificación utilizados, no dando lugar a reasignaciones de empresas entre los grupos. Por otra parte, en todos los casos el ajuste de cada empresa a su grupo fue muy bueno.

El cálculo de la distancia de Mahalanobis para estos tipos resultó en:

$$D^2 = 1856.4$$

El valor tabulado correspondiente (al 1%) es:

$$x^2_{60.01} = 88.4$$

Lo que daría gran confianza en cuanto a la independencia estadística entre los grupos conformados de manera subjetiva.

Esta impresión se reafirma al observarse la porción del output correspondiente a la evaluación de las funciones de clasificación para cada observación. Efectivamente: todos los casos están bien clasificados por su correspondiente función. Además las probabilidades oscilan entre .99 y 1 (con sólo un caso de .94! !).

Sobre las posibilidades del Análisis Discriminante, podemos recordar que:

1. Puede ser útil para comprobar agrupaciones realizadas, fundamentalmente cuando los criterios utilizados generan dudas en cuanto a su poder clasificador.
2. También presenta utilidad para asignar nuevos casos que puedan a priori pertenecer a más de un grupo conformado.
3. No genera grupos óptimos ante cualquier criterio, sino que hay numerosas agrupaciones posibles y por tanto los criterios de agrupación deben ser definidos anteriormente.
4. El análisis discriminante comprobará el poder de esos criterios y su correcta aplicación en cada caso.

En el caso de la clasificación realizada sobre los tambos encuestados, el criterio utilizado para agrupar dio resultados muy atractivos. Sin embargo, esto último no significa que la validez estadística sea rotunda ni que necesariamente ese criterio sea el más adecuado a los fines propuestos.

Para afirmar la validez estadística de los resultados (que aparecen como muy buenos, según se reportó) habría que tener mayor seguridad de que el número de variables no era demasiado grande en relación al tamaño de la muestra, que resultó incluso en un grupo con sólo cinco casos*.

* Sobre el mismo asunto, véase el trabajo de Dabezies y Sarroca presentado a esta Reunión.

Pero, aún dada la validez estadística, nada prueba que la clasificación retenida sirva realmente a los propósitos del estudio. A este efecto, tal vez hubiera sido mejor partir de una hipótesis más precisa sobre los factores que determinan diferencia tecnológica. Sobre esa base pudo haberse aplicado alguna técnica de conglomeración con un número reducido de variables. La prueba de diferencia entre conglomerados, pudo haber pasado por docimar la estabilidad de los mismos ante métodos alternativos* y terminado con aplicación de Discriminante en función de algunas variables resultado (tales como producción de leche por hectárea).

Dado que el test de Análisis Discriminante empleado no satisfizo plenamente el propósito de validación objetiva, porque no fue insertado en un proceso convincente de tipificación - validación, se decidió retenerlo y presentarlo sólo como parte de un esquema de aprendizaje. Y, con respecto a la tipificación de los tambos, se consideró oportuno reconsiderarla desde su comienzo.

7.7 Correlaciones entre variables

Al comenzar el uso de técnicas estadísticas, fácilmente se abre el interés de avanzar por el camino que ellas ofrecen para mejor explorar el universo en estudio, mejor formular y probar hipótesis sobre el mismo.

La tipificación realizada con los predios pasó con holgura la prueba de Análisis Discriminante, lo que, no obstante ello, se consideró una débil prueba de objetividad. Por esto se recurrió a emplear información contenida en el output de Componentes Principales, como forma de recomenzar el proceso de agrupamiento de tambos.

Primero, y siguiendo sugerencias del Prof. Marsal, se trabajó con la matriz de correlación entre las 20 variables de que se disponía para los predios lecheros. La idea con que se realizó este estudio consistió en ver, de manera elemental, que decía el universo de datos en cuanto a relaciones de las variables entre sí. Del análisis de esta información ya podría deducirse algo en cuanto a la estructura de los predios; algo, en definitiva, que aportara a todo el proyecto que requirió estudiarlos, incluyendo sugerencias sobre cómo agruparlos.

A estos efectos, la matriz de correlaciones se reprodujo gráficamente a distintos niveles de significación estadística preimpuesta. Para un dado nivel de significación, el gráfico sólo retiene símbolos (+ y –) que indican la dirección de una relación estadísticamente no rechazable, dejando el resto de los casilleros de la matriz en blanco.

Para estos gráficos pueden ordenarse las variables, de manera que queden submatrices con conjuntos de variables relacionadas.

A medida que aumente la tolerancia estadística, es esperable que:

a) las submatrices aumenten de tamaño, y

b) se incrementen las áreas de intersección entre submatrices

A modo de ejemplo, se acompaña a este informe la gráfica del Cuadro No. 3, resultante del nivel 1%. Adviértase que se requiere una correlación como mínimo para estar en un grupo no trivial.

En el caso graficado se generan tres grupos definidos de variables, tres variables aisladas y un área de intersección entre grupos (variables: lts/vaca y Hás/Potrero).

---

* Para la propuesta de verificar estabilidad de los conglomerados, véase la propuesta de P. Ferreira presentada a esta Reunión.

Cuadro 3

Matríz de correlaciones — Nivel 1%

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1.	Lts/hectárea		+	−	+	+	−														
2.	% Praderas Permanentes	+		−	+	+															
3.	1er. Entore (meses)	−	−		−																
4.	Carga	+	+	−																	
5.	Lts/vaca	+	+											+	+						
6.	Hás/Potrero	−							+	+	+				+						
7.	% Unidad lechera Vacas/U. L. Total									−					−						
8.	Superficie Total						+			+	+	+		+	+						
9.	Superficie Lechera						+	−	+		+			+	+						
10.	Vacas Masa						+		+	+				+	+						
11.	% Cultivos Comerciales								+				+								
12.	% Praderas Nuevas											+		+							
13.	Número Potreros					+			+	+	+		+		+						
14.	Cuota					+	+	−		+	+			+							
15.	% Alfalfa																	+			
16.	Cuota/Hectárea																	+			
17.	Ración (grs./litro leche)															+	+				
18.	Vacas Producción/Vacas Secas																				
19.	% Cultivos Anuales																				
20.	% Praderas Viejas																				

Realizado el estudio de la matriz de correlación a 4 niveles de significación se puede establecer grupos de variables asociadas, a la vez de definirse variables no correlacionadas entre sí.

A niveles de significación muy altos ya aparecen dos grupos, uno asociado a la intensidad de la explotación y el otro al tamaño (o dimensión) de la empresa, compuestos por tres o cinco variables respectivamente.

Disminuyendo el nivel de significación al 1%, aparece un tercer grupo de variables asociadas entre sí. Este está constituído por 3 variables vinculadas a la proporción de cuota del predio, cantidad de ración utilizada y el porcentaje de alfalfa. Los grupos anteriores se amplían a 6 y 10 variables (intensidad y dimensión respectivamente) apareciendo 2 de ellas vinculadas a ambos aspectos (lts/vaca/día y tamaño de potreros).

Si se continúa el proceso, los grupos continuarán ampliándose en cuanto al número de variables que intervienen en cada uno y aumentando las que participan en más de uno.

Como interpretación de este análisis se pueden extraer las siguientes conclusiones:

1. La intensidad de producción de leche está asociada fundamentalmente al porcentaje de praderas, carga, primer entore temprano, todo lo que lleva a altas producciones por hectárea. La producción por vaca por día y el menor tamaño de potreros también inciden, aunque en menor grado, sobre este aspecto.
2. La intensidad de producción está poco vinculada a variables de dimensión de los establecimientos, salvo la relación inversa con hectáreas/potrero.
3. A la dimensión, aparte de las variables claramente de tamaño (superficie total, superficie lechera, número de vacas masa, total de leche cuota), se asocian potreros de mayor tamaño, un menor porcentaje de vacas en el rodeo (que indica una tendencia a la cría de reemplazos), obtención de mayores producciones por vaca por día (probablemente vinculado a una menor carga y la flexibilidad de manejo que puede dar el mayor porcentaje de reemplazos). Existiría también tendencia a los establecimientos mayores a la realización de cultivos comerciales y a la instalación de praderas permanentes. Esta última tendencia tal vez sea de carácter reciente, ya que no existe correlación del porcentaje de praderas permanentes en producción con el tamaño.
4. El tercer grupo de variables correlacionadas se hace más difícil de identificar con algún aspecto específico de la explotación. Las variables de este grupo pueden determinar la "forma" de producción del establecimiento a través de una alta proporción de cuota y el uso de mayores cantidades de concentrados. Junto a esto, aumenta el porcentaje de alfalfa en los predios, el número de potreros y la proporción vacas en ordeñe a vacas secas en invierno, todo ello sin incidir en la producción de leche por hectárea.
5. Finalmente el porcentaje de cultivos anuales, así como el porcentaje de praderas viejas, aparecen desvinculadas totalmente de los tres grupos (o aspectos) detectados.
6. Analizando las vinculaciones de algunas variables resulta:
   a. La producción por há. aparece asociada fundamentalmente al porcentaje de praderas permanentes en producción y a la carga. También se vincula a entores a baja edad de las vaquillonas, mayores producciones diarias por vaca, alto porcentaje de vacas en el total del rodeo (se puede vincular a la menor edad en que los reemplazos se integran a la producción o a una mayor especialización lechera criando sólo los necesarios) así como potreros de menor tamaño. Es importante destacar que algunas variables de aparente importancia en la producción no presentan correlación con el resultado por há., siendo aquellas la utilización de concentrados, los cultivos forrajeros anuales y la alfalfa. Tampoco la dimensión de la explotación resulta muy claramente en una mayor extensividad de la producción, aunque hay algunos indicios de ello.

b. El porcentaje de praderas permanentes en producción se vincula estrechamente, como ya fue dicho, a la variable anterior, así como a la edad de primer entore, la carga y la producción por vaca, y también al menor tamaño de potreros y a la menor utilización de ración.

c. Interesa destacar que la producción por vaca por día se encuentra asociada no sólo a la intensidad de la explotación, sino también a la dimensión de ésta (cuota total y número de potreros).

d. El número de potreros aparece asociado a la dimensión, fundamentalmente, a la producción por vaca y a la "forma" de producción (3er. grupo de variables). No tiene correlación con la intensidad de producción, como sería de esperar. Sin embargo, el tamaño de los potreros sí aparece con dicha vinculación.

e. La utilización de ración está relacionada con el nivel de cuota (cuota/há.) que tiene el establecimiento y al porcentaje de alfalfa. También lo está, en menor grado, a la relación vaca en ordeñe a vacas secas en invierno y negativamente a la proporción de praderas. De aquí se puede concluir, que, si bien la ración no incide en la producción total de leche, permite mantener mayores niveles de cuota. La existencia de niveles altos de praderas permite, a su vez, ahorros en el uso de concentrados.

f. Es importante notar la desvinculación existente entre el porcentaje de cultivos anuales y el nivel de producción (tanto por hectárea como por vaca) y con el nivel de carga del predio y la cuota por hectárea. De aquí se puede inferir o bien que este porcentaje es un factor fijo de la producción o bien que no incide sobre ella. La tercera posibilidad a considerar es que no todos los cultivos anuales tienen el mismo comportamiento, lo que haría necesario desglosar esta variable en sus distintos componentes según época de producción y/o especies. Esto fue hecho con los cultivos permanentes (praderas en producción, nuevas, viejas y alfalfa) resultando en distintos comportamientos para cada uno.

Desde el punto de vista tecnológico, tenemos existencia de dos grupos de variables no correlacionadas: uno asociado al nivel de producción y otro a la continuidad de la producción (cuota) y que la dimensión de la explotación tiene poca incidencia sobre la tecnología utilizada, aunque se percibe alguna tendencia a una relación inversa.

A los efectos de la evaluación de los niveles técnicos empleados, podrían utilizarse agrupaciones de establecimientos en base a altos o bajos valores para cada grupo de variables o combinaciones de ambos. Se dispone así de una serie de elementos de juicio orientadores sobre el comportamiento de variables que aparecía confuso o era difícil de cuantificar a partir del planillado de datos original.

7.8 Principales componentes

Siguiendo con el intento de mejor explorar el contenido de los datos, se procedió a revisar todo el output típico de análisis factorial en base a componentes.

El resultado, presentado para las 4 primeros componentes, y con valores elevados al cuadrado, se incluye en el Cuadro No. 4.

A juicio del equipo de trabajo, resultaron dos primeros factores de muy clara interpretación, siendo del tercero en adelante difícil de visualizar el significado en forma precisa.

El primer factor representa fundamentalmente la dimensión del predio, explicando entre un 80 y 94% de la varianza de variables de tamaño: superficie total, superficie lechera, cuota total, número de vacas masa. En menor grado, se vincula con el tamaño de los potreros. Aparece luego explicando un 20% o menos de algunas variables tecnológicas, pudiendo esto representar que predios mayores realizan prácticas de producción algo más extensivas y son algo más diversificados que los chicos, teniendo menor porcentaje de vacas en el rodeo (más cría de reemplazos) y mayor porcentaje de tierra dedicada a cultivos comerciales.

La dimensión (por varias medidas) no se asocia con el nivel de cuota por hectárea ni con la intensidad de utilización de concentrados y cultivos anuales en la explotación.

El factor que aparece en segundo término está íntimamente vinculado a la tecnología o a la "intensidad de la explotación", explicando cerca del 65% de la variación en producción de leche por há. y en el porcentaje de praderas permanentes. A través de este factor se explica también entre un 25 y un 50% de la variación de otras variables tecnológicas: litros por vaca por día, carga, edad de primer entore y número de potreros. La intensidad de producción está también vinculada, aunque en un bajo grado, a un menor uso de concentrados, potreros más chicos y mayor número de vacas. Se debe hacer notar la inexistencia de asociación de este factor con los porcentajes de cultivos anuales y alfalfa, la cuota por hectárea y la relación vacas en ordeñe a secas en invierno. Estas variables evidencian, entonces, poca incidencia en la "intensidad" de producción.

Estos dos primeros factores explican el 25% y 16%, respectivamente, de la variación total en la información aportada.

Cuadro 4

Matríz variables — Componentes sin rotar
Cuadrados de las correlaciones entre variables y factores (varianzas)*
Primeros tres factores

		I	II	III	
	Factor:				
	Aporte porcentual del Factor a la varianza total.	25.2	16	10.6	Σ = 51.8%
	Variables:				
1.	Lts/hectáreas	11 (–)	66		
2.	Porcentaje de Praderas Permanentes	11 (–)	64		
3.	Edad Primer Entore		32 (–)		
4.	Carga	19 (–)	36		
5.	Lts/vaca		50		
6.	Hás/Potrero	47		19 (–)	
7.	Porcentaje Vacas sobre Rodeo (Unidades lecheras)	22 (–)			
8.	Superficie Total	85			
9.	Superficie Leche	94			
10.	Vacas Masa	81			
11.	Porcentaje Cultivos Comerciales	11			
12.	Porcentaje de Praderas Nuevas			13	
13.	Número de Potreros	15	26	22	
14.	Cuota Total	84			
15.	Porcentaje Alfalfa			26	
16.	Cuota/Hectárea			26	
17.	Ración		10 (–)	50	
18.	Vacas Producción/Vacas Secas			28	
19.	Porcentaje Cultivos Anuales			20	
20.	Porcentaje de Praderas Viejas				

* Para mejor visualización, se excluyen aportes inferiores al 10% (correlaciones inferiores en valor absoluto a 3.16). Los signos de las correlaciones originales negativas se indican entre paréntesis, a la derecha de las varianzas.

En tercer término surge un factor cuyas variables preponderantes ya no están asociadas en tan alto grado como en los anteriores. La variable ración es la de más fuerte vinculación al factor, el que explica un 50% de su variación. A su vez explica entre un 28 y 19% de la cuota por há., relación vacas en ordeñe a vacas secas, porcentaje de alfalfa y de cultivos anuales, tamaño y número de potreros. Se puede definir este factor como tendencia a la cuota, en base a una cierta organización del rodeo y, fundamentalmente, alimentación con concentrados, cultivos anuales y alfalfa, con alguna inclinación a la implantación de praderas. Podría tal vez denominarse "orientación" de la lechería, aunque ya se entra acá en terreno de interpretaciones no muy claras.

El cuarto factor es difícil de interpretar, siendo la variable preponderante el porcentaje de praderas nuevas, siguiendo en importancia el porcentaje de cultivos comerciales. Está asociado también a mayor porcentaje de vacas en el total del rodeo, menos lts/vaca/día, menor cuota por há. y baja carga. Puede inferirse de aquí una tendencia de los establecimientos agrícola-lecheros a la incorporación de praderas y a una forma de producción menos intensiva, que no incidiría en la producción por há., en base a cierta estacionalidad (menos cuota por há.). El factor no permite sin embargo ser concluyente en cuanto a su significación, debido a la diversidad del carácter de las variables que se le asocian.

Los restantes factores no se vinculan a aspectos definidos. A efectos de mejor interpretar esta información, puede realizarse un análisis de algunas variables, como sigue.

La producción por há. se halla fuertemente vinculada con la intensidad lechera (66%), definida esta por el uso de altos porcentajes de praderas, mayor carga, más potreros. La dimensión del predio tiene a su vez una relación negativa, pero poco importante, con dicha producción; una cierta tendencia a relación inversa productividad - tamaño sigue presente.

El porcentaje de praderas presenta un comportamiento similar, un 64% de su variación se vincula a una mayor producción (lts. por há., lts. por vaca, carga, etc.) mientras que en un 11% se debe a la dimensión del predio, con tendencia de los mayores a hacer algo menos de praderas, en porcentaje, que los más chicos.

La producción por vaca por día se vincula a la intensidad lechera (50%).

La utilización de cultivos anuales puede plantearse como independiente de la dimensión e intensidad del predio lechero y se vincula más bien a mayor cuota y uso de ración (orientación lechera).

El consumo de ración por litro de leche se explica en primer lugar por el factor de orientación lechera (50%) interpretándose que, o bien permite mantener niveles relativamente altos de cuota/há., o bien se asocia a la organización de la explotación, sin incidir en la producción por hectárea.

En nuestro caso la rotación (o reajuste) de los factores no aportó mayor esclarecimiento sobre el carácter de los dos primeros factores (dimensión e intensidad lechera) y volvió más oscura la interpretación de los siguientes. En todo caso, sirvió para confirmar la existencia de sólo dos factores bien definidos.

Creemos que este tipo de análisis realmente permite comprender la estructura implícita en la masa de datos generada por la encuesta. En el próximo capítulo se propone un posible uso de las conclusiones obtenidas, para el propósito de repensar todo el problema.

7.9 Posibles replanteos del caso en base al análisis factorial

Hecha la revisión de los tipos obtenidos en la clasificación subjetiva inicial, se extraen las siguientes conclusiones:

1. El nivel tecnológico (o "intensidad lechera") podría plantearse como independiente de la dimensión del predio, con cierta pérdida de información. Esto requeriría llevar todos los predios tipo

a un tamaño uniforme. Se resolvería así un problema originado en diferencias en el resultado económico de los predios que se deben exclusivamente a la dimensión de la empresa, para un dado nivel de tecnología utilizada. Como se recordará, el objetivo de todo el proyecto es el de comprobar la viabilidad económica de tecnologías intensivas. No obstante ello, parecería mejor reclasificar con una técnica de conglomeración que, registrando algún efecto de dimensión, permitiera un mejor análisis posterior de los elementos que determinan el nivel tecnológico.

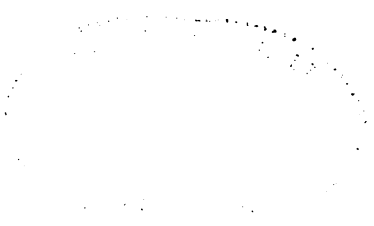
2. En los tipos formulados resultaron variar en la misma dirección (como se advierte en el Cuadro No. 2), características de "intensidad" (litros por hectáreas, porcentaje de praderas y carga) y características de "orientación lechera" (cuota/há., porcentaje de alfalfa). La "orientación lechera", no obstante, aparece como un posible tercer factor latente, con variables propias.

Para tomar debidamente en cuenta estas últimas, se deberían establecer dentro de los tipos de "intensidad", subtipos con distinta "orientación". También ésto podría dilucidarse con alguna técnica jerárquica de conglomeración y su análisis posterior.

3. Fue razonable la simplificación realizada al no incluirse cultivos comerciales en los tipos estructurados, ya que estos cultivos no parecen tener vinculación con las características bajo análisis, salvo que se relacionan con el tamaño del predio.

CAPITULO 8

Subregionalización mediante análisis de conglomerados.



Subreglonalización mediante análisis de conglomerados.

8

Alfredo Alonso – DIEA

8.1 Introducción

En este trabajo se presenta una agrupación de unidades administrativas como aporte al Proyecto de Desarrollo Regional en Uruguay (convenio MAP - IICA).

La aplicación está orientada a la definición de subregiones dentro del departamento de Cerro Largo, que forma parte junto con Tacuarembó y Rivera de la región Plan del Proyecto.

A partir de la poca experiencia acumulada en la aplicación de técnicas estadísticas de clasificación a la tipificación de empresas agropecuarias (o distritos censales, en este caso), nos hemos planteado una metodología que se podría resumir en tres etapas consecutivas. Ellas son:

i. Exploración de las variables o atributos
ii. Clasificación o conglomeración.
iii. Análisis de agrupamientos obteñidos.

Este trabajo se concentra sobre la segunda etapa, presentando una aplicación de dos métodos de clasificación disponibles en el IICA - Uruguay. El método de "Ward", que en forma más bien intuitiva aparece como muy indicado para clasificar empresas en un proceso de tipificación, se compara con el método de "Vecino más cercano", que opera en forma diferente y que no aparece como recomendable para los casos más generales de tipificación*. En la medida en que un estudio de regionalización, al igual que la tipificación de empresas, resulta en definitiva en un problema de clasificación, se trata de probar la aplicabilidad de estos métodos a trabajos de regionalización.

8.2 Variables utilizadas

La información utilizada proviene del Censo General Agropecuario de 1970 tabulada a nivel de sector censal. El departamento de Cerro Largo se encuentra dividido en 47 sectores censales. Estos sectores constituyen los elementos a clasificar, como paso previo a la definición de subregiones.

* Para un mayor detalle sobre estos métodos, ver el Capítulo 3 de esta publicación.

Los sectores censales no son unidades naturales sino que se han diseñado por subdivisión de las secciones policiales de acuerdo con necesidades y objetivos censales. Dado que estos sectores difieren sustancialmente en tamaño, no se han considerado variables en términos absolutos, prefiriéndose la utilización de relaciones entre atributos para tratar de caracterizarlos.

Las variables han sido seleccionadas de acuerdo con las características propias del departamento considerado. Para el departamento de Cerro Largo fueron censadas en 1970 un total de 1:311.860 hectáreas (7,94% del país) y 4.156 explotaciones agropecuarias (5,39% del país). Se trata de un departamento fundamentalmente ganadero, donde las tierras dedicadas a agricultura representan sólo un 2,32% del total, frente a un 7,73% para el país.

Los atributos considerados se pueden agrupar en dos clases:

i. la primera, constituída por 6 variables, trata de identificar algunas características generales de los sectores, como ser: tamaño promedio de las explotaciones, intensidad de mano de obra, mano de obra familiar, grado de mecanización, tamaño promedio de los potreros, y mejoramiento de pasturas naturales.

ii. la segunda, constituída por 3 variables, trata de identificar los rubros predominantes: agricultura o ganadería y (dentro de esta última) producción ovina, bovina o de leche.

Las variables empleadas fueron:

1. **Tamaño promedio de las explotaciones:**
   Superficie censada en cada sector censal dividida por el número de explotaciones correspondiente.

2. **Hectáreas por trabajador:**
   Hectáreas censadas divididas por el número de trabajadores de 14 años y más.

3. **Mano de obra familiar:**
   Número de trabajadores de 14 años y más comprendidos en la categoría "productor y miembros de su familia" dividido por el número total de trabajadores de 14 años y más.

4. **Hectáreas por tractor:**
   Número total de hectáreas censadas dividido por el número total de tractores.

5. **Tamaño de los potreros:**
   Superficie total censada dividida por el número total de potreros.

6. **Relación Agrícola - Ganadera**
   Superficie dedicada a agricultura dividida por la superficie dedicada a ganadería (en porcentaje)

7. **Superficie mejorada:**
   Superficie mejorada (campo natural fertilizado, sembrado a zapatas, sembrado en cobertura y praderas artificiales permanentes) dividida por la superficie dedicada a ganadería (en porcentaje).

8. **Relación Ovino - Bovino:**
   Número total de ovinos dividido por el número total de bovinos.

9. **Ganado lechero:**
   Total de ganado lechero sobre el total de vacunos (en porcentaje).

En el Anexo se encuentra la matriz de observaciones y la matriz normalizada.

8.3 Métodos de clasificación utilizados

Para agrupar los sectores censales se utilizó el algoritmo de Wishart* aplicando el método de "Ward" que permite obtener una estimación de la varianza dentro de los grupos resultantes y el método de "Vecino más Cercano".

En ambas aplicaciones se estandarizaron las variables para obviar el problema de las unidades de medida y no se encontró razón para ponderar los atributos.

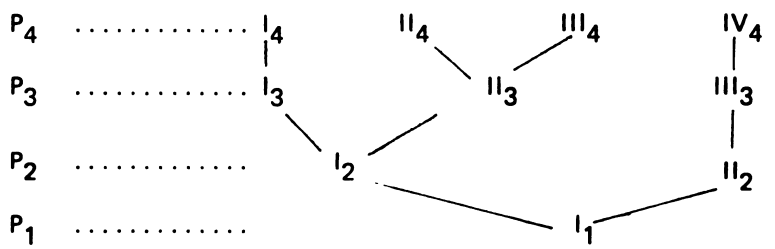
8.4 Resultados obtenidos

A. Metodo de Ward

En el cuadro No. 1 se presentan los agrupamientos resultantes para las particiones en 3 y 4 clusters. Para 5 clusters aparecen conglomerados que contienen un sólo sector censal, de modo que esta partición no aporta mayor información al problema planteado.

En la figura No. 1 se esquematiza la forma en que se van uniendo los clusters desde la partición en 4 conglomerados hasta llegar a la partición monotética.

Figura 1



* Alonso, A. "Algunas técnicas de conglomeración . . .", capítulo 3 del presente volumen.

Cuadro 1

Conglomerados resultantes de la aplicación del método de Ward

CLUSTER	NUMERO DE SECTORES	SECTORES CENSALES
I_4	12	3 - 4, 3 - 5, 5 - 5, 5 - 6, 5 - 7, 5 - 8, 11 - 1, 11 - 2, 14 - 1, 14 - 2, 14 - 3, 14 - 4
II_4	9	3 - 1, 3 - 2, 3 - 3, 10 - 1, 12 - 2, 13 -1, 13 - 2, 13 - 5, 13 - 6
III_4	19	4 - 1, 4 - 2, 4 - 3, 4 - 4, 5 - 4, 6 - 2, 7 - 1, 7 - 2, 8 - 2, 8 - 3, 9 - 2, 9 - 3, 10 - 2, 10 - 3, 12 - 1, 12 - 3, 13 - 3, 13 - 4, 14 - 5
IV_4	7	5 - 1, 5 - 2, 5 - 3, 6 - 1, 8 - 1, 9 - 1, 10 - 4

CLUSTER	NUMERO DE SECTORES	SECTORES CENSALES
I_3	12	3 - 4, 3 - 5, 5 - 5, 5 - 6, 5 - 7, 5 - 8, 11 - 1, 11 - 2, 14 - 1, 14 - 2, 14 - 3, 14 - 4
II_3	28	3 - 1, 3 - 2, 3 - 3, 4 - 1, 4 - 2, 4 - 3, 4 - 4, 5 - 4, 6 - 2, 7 - 1, 7 - 2, 8 - 2, 8 - 3, 9 - 2, 9 - 3, 10 - 1, 10 - 2, 10 - 3, 12 - 1, 12 - 2, 12 - 3, 13 - 1, 13 - 2, 13 - 3, 13 - 4, 13 - 5, 13 - 6, 14 - 5
III_3	7	5 - 1, 5 - 2, 5 - 3, 6 -1, 8 - 1, 9 - 1, 10 - 4

En la figura No. 2 se grafica la suma de cuadrados dentro de los clusters, como viene dada por la función objetivo, con respecto al número de conglomerados.

En el cuadro No. 2 aparece el valor de la función objetivo correspondiente a cada partición y el incremento que se produce cada vez que se unen dos conglomerados.

Al disminuir el número de clusters, estos se van haciendo cada vez menos homogéneos, aumenta la intravarianza y disminuye la varianza entre conglomerados.

Se puede apreciar que los incrementos que se van produciendo en la función objetivo crecen lentamente hasta llegar a la partición en 3 clusters y que a partir de ésta el valor de la intravarianza sube rápida-

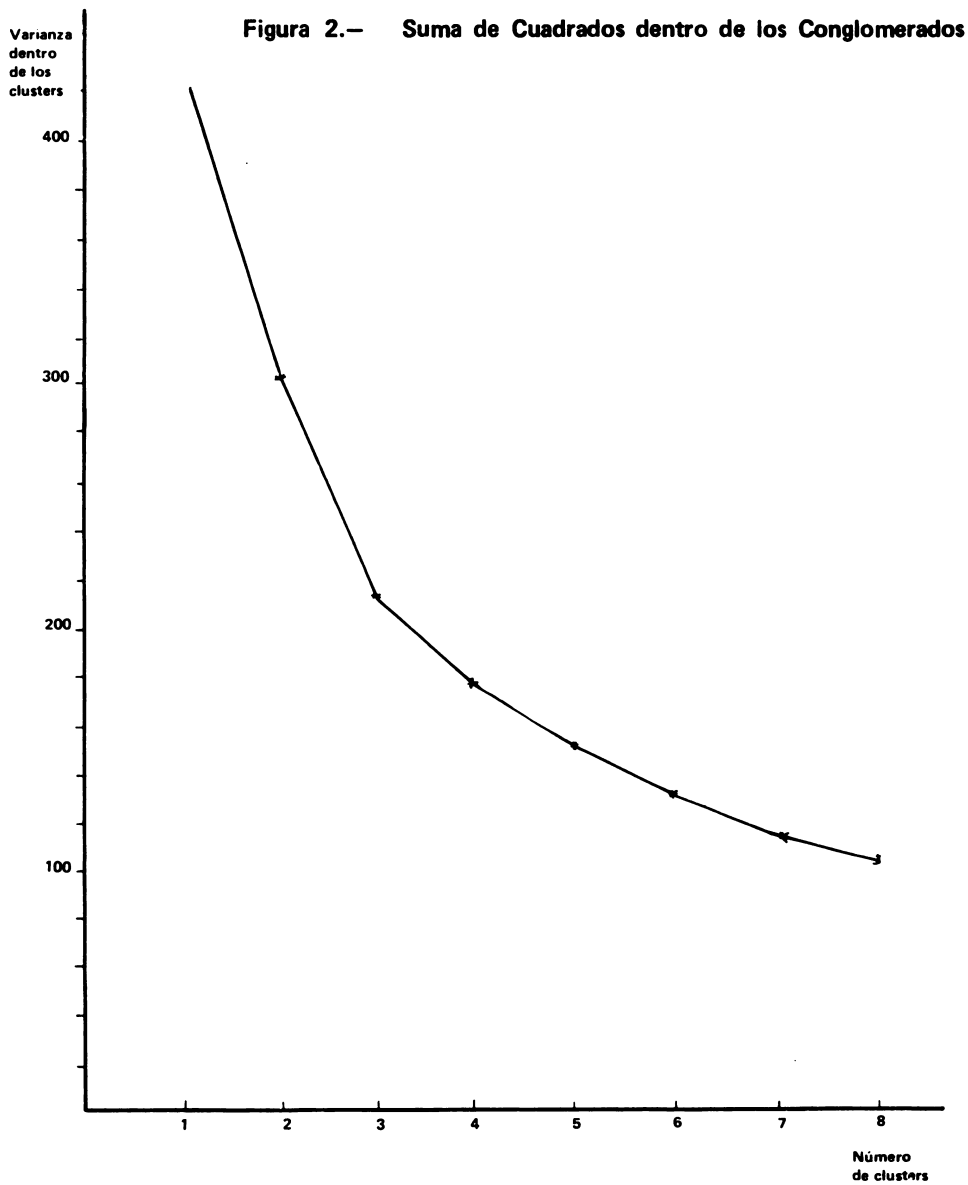
mente. Esto permite afirmar que, dada la naturaleza de los datos, no se pueden elegir menos de 3 clusters, si se pretende trabajar con conglomerados razonablemente homogéneos.

En el cuadro No. 3 se ve que la varianza dentro de los conglomerados para la partición en 4 clusters representa el 42% de la varianza total, de modo que la varianza entre los clusters es el 58% de la total. Con 3 clusters se tiene la varianza total dividida en partes iguales entre los conglomerados y dentro de los conglomerados.

De acuerdo con esto se puede considerar adecuada la partición en 4 clusters y a partir de ella se tratarán de definir las subregiones del departamento considerado.

Para facilitar la interpretación de los resultados se presentan en el cuadro No. 4 los valores medios de las variables consideradas para los clusters que surgen en las dos particiones (P_3 y P_4), junto con los valores medios correspondientes al departamento.

Figura 2.– Suma de Cuadrados dentro de los Conglomerados



Cuadro 2

Función Objetivo, Valor Absoluto e Incremento para cada Partición

NUMERO DE CLUSTERS	VALOR DE LA FUNCION OBJETIVO	INCREMENTO DE LA FUNCION OBJETIVO (%)
7	116,21	——
6	130,84	12,59
5	151,96	16,14
4	178,51	17,47
3	212,27	18,91
2	301,64	42,10
1	422,99	40,23

Cuadro 3

Descomposición de la Varianza Total

NUMERO DE CLUSTERS	VARIANZA DENTRO DE LOS CLUSTERS (%)	VARIANZA ENTRE LOS CLUSTERS (%)
4	42	58
3	50	50
2	71	29
1	100	—

Cuadro 4

Valores Promedio Correspondientes a los Conglomerados

Cluster	Tamaño Promedio	Hectáreas por Trabajador	Mano de Obra Familiar	Hectáreas por Tractor	Tamaño Potreros	Relación Agrícola - Ganadera	Superficie Mejorada	Relación Ovino - Bovino	Ganado Lechero
I_4	154.17	69.39	0.78	968.42	45.36	8.27	8.19	1.24	7.42
II_4	181.87	74.32	0.88	1.911.28	64.87	2.56	3.97	2.79	1.25
III_4	420.94	189.34	0.69	3.464.57	105.00	1.20	5.33	1.92	0.70
IV_4	1.736.07	342.15	0.32	2.871.63	201.26	0.88	3.96	1.45	0.68
I_3	154.17	69.39	0.78	968.42	45.36	8.27	8.19	1.24	7.42
II_3	344.09	152.37	0.75	2.965.30	92.10	1.63	4.89	2.20	0.88
III_3	1.736.07	342.15	0.32	2.871.63	201.26	0.88	3.96	1.45	0.68
Departamento de Cerro Largo	315.65	131.66	0.72	1.824.56	87.43	2.43	5.29	1.73	1.84

La **subregión I_4** está formada por "sectores representados por explotaciones chicas, tecnificadas, de producción agrícola - lechera". La dotación de mano de obra por hectárea es alta, están muy mecanizadas y el tamaño de los potreros es reducido.

La **subregión II_4** está formada por "sectores representados por explotaciones chicas, familiares, de producción ovina". El nivel de mecanización es medio y la dotación de mano de obra es alta y de tipo familiar.

La **subregión III_4** está constituída por "sectores representados por explotaciones medianas, poco tecnificadas de producción ganadera". La mano de obra es baja, muy poco mecanizadas y con una proporción media de pasturas mejoradas.

La **subregión IV_4** está constituída por "sectores representados por explotaciones muy grandes, extensivas de producción bovina". Usan muy poca mano de obra, que es asalariada, y el tamaño de los potreros es elevado.

Si se analiza la partición en 3 clusters se tiene que el nuevo conglomerado II_3 formado por la unión de II_4 y III_4 podría caracterizar a una subregión formada por "sectores representados por explotaciones medianas, poco tecnificadas de producción ovina".

B. Método de Vecino más Cercano

Al aplicar este método, los sectores tienden a unirse a clusters ya formados en lugar de convertirse en centros de nuevos clusters. Se forma así un gran conglomerado que contiene a la mayoría de los sectores, quedando sectores aislados que por alguna razón se diferencian netamente del resto.

Si se analiza, por ejemplo, la partición en 4 conglomerados que aparece en el cuadro No. 5 se tienen 44 sectores en un conglomerado restando 3 clusters singulares.

Cuadro 5

Conglomerados Resultantes de la Aplicación del Método de Vecino más Cercano

CLUSTER	NUMERO DE SECTORES	SECTORES CENSALES
I	1	3 - 4
II	1	13 - 4
III	1	14 - 3
IV	44	El resto

Analizando la matriz de observaciones, se puede ver que los sectores que forman conglomerados singulares presentan valores extremadamente altos, con respecto al promedio general, para alguna de las variables consideradas.

Los sectores 3 - 4 y 13 - 4 presentan un porcentaje de superficie mejorada de 19.81 y 15.39 respectivamente, frente al promedio departamental que es de 5.29. El sector 14 - 3 presenta un porcentaje de ganado lechero de 27.38 frente a un promedio departamental de 1.84.

Dada su conformación, se puede concluir que los conglomerados obtenidos no sirven para definir sub - regiones.

De acuerdo con los objetivos más generales de regionalización o tipificación, este método no aparece como recomendable. Sin embargo, enfrentados a otro tipo de problemas, puede resultar de mucha utilidad Supongamos que se plantea una política de control de erosión de los suelos. Se pueden caracterizar las regiones de acuerdo con un grupo de variables que aporten información sobre el grado actual de erosión y características de los suelos, uso de la tierra y otros atributos que den idea sobre el riesgo potencial de erosión. Aplicando este método, se podrían aislar las regiones que se encuentran en una situación más crítica de acuerdo con el problema planteado y que deben ser asistidas en forma prioritaria.

8.5 Concordancia de las clasificaciones obtenidas

La utilidad de las clasificaciones se debe plantear en la posibilidad que ofrezcan de ejemplificar algún principio de agrupamiento que sea coherente con los objetivos del trabajo planteado. Para ello puede estudiarse la concordancia que presenten los agrupamientos obtenidos con respecto a algún criterio o variable que resulte relevante en el contexto general.

En este ejemplo se plantean tablas de contingencia para docimar la existencia de asociación entre el hecho de que los sectores pertenecen a un conglomerado determinado y el valor que presentan para la variable: tamaño promedio de las explotaciones.

En una segunda etapa se trata de medir el grado de dependencia entre los criterios planteados mediante el estadígrafo de contingencia "C" de Pearson.

Al analizar los resultados se deben tener presente las reservas que surgen del hecho de trabajar con un número reducido de observaciones, lo que determina que aparezcan muchas celdillas vacías.

En el Cuadro No. 6 se plantea una tabla de contingencia comparando los agrupamientos con el valor que presentan para la variable tamaño promedio de las explotaciones en cuatro tramos: menos de 200 hectáreas, de 200 a 500, de 500 a 1.000 y más de 1.000 hectáreas.

Al comparar el valor de T calculado con el valor de X^2 (con 9 grados de libertad a nivel 0.01), se rechaza la hipótesis nula de independencia entre las características consideradas.

Al medir el grado de asociación mediante el estadígrafo "C" de Pearson se observa que el valor calculado es alto comparado con el máximo que puede tomar para una tabla de contingencia de estas características.

De acuerdo con los resultados obtenidos se podría concluir que la clasificación obtenida es concordante con la variable analizada; sin considerar otras pruebas de validación que deben ser realizadas, y aceptando que esta variable es relevante dentro del contexto general del problema planteado, la clasificación obtenida puede ser usada con un cierto grado de confianza para definir subregiones homogéneas dentro del universo considerado.

Cuadro 6

Tabla de Contingencia

Superficie \ Clusters	I	II	III	IV	f. m.
– 200	10 (4.60)	7 (3.45)	1 (7.28)	0 (2.68)	18
200 - 500	2 (4.34)	2 (3.26)	13 (6.87)	0 (2.53)	17
500 - 1000	0 (1.53)	0 (1.15)	5 (2.43)	1 (0.89)	6
+ 1000	0 (1.53)	0 (1.15)	0 (2.43)	6 (0.89)	6
f. m.	12	9	19	7	47

$$T = \Sigma \frac{(F_o - F_t)^2}{F_t}$$

$T = 67{,}699$

$X^2_{(9,0.01)} = 21.666$

$$C = \sqrt{\frac{T}{T + N}}$$

$C = 0.768$

máx. $C = 0.866$

8.6 Conclusiones generales

Resulta difícil, en el estado actual del conocimiento que se posee sobre aplicaciones de técnicas del Análisis de Conglomeración a la tipificación o regionalización, el aconsejar el uso de determinados métodos.

De acuerdo con los objetivos específicos de los trabajos que se planteen y la naturaleza de los datos, los distintos métodos pueden presentar ventajas relativas que podrán evaluarse al realizar las aplicaciones concretas.

En forma general se puede decir que el método de "Ward" puede ser utilizado con éxito en procesos de regionalización, debido a que los agrupamientos obtenidos parecen adecuados para servir de base a la definición de regiones homogéneas.

El método de "Vecino más cercano" no permite obtener conglomerados que puedan ser utilizados para definir regiones o empresas tipo en el sentido más general. Sin embargo, si se deseara seleccionar regiones que sean marcadamente diferentes del resto, de acuerdo con algún atributo o grupo de atributos dado, este método sería el más aconsejable.

De cualquier manera, al plantear un trabajo de regionalización en varias etapas partiendo de una regionalización a nivel global, para luego ir afinando a nivel de subregiones, se puede plantear la utilización de distintos métodos.

En cada etapa tanto las variables como los métodos de clasificación pueden variar sustancialmente en la medida en que cambia la naturaleza de los elementos a agrupar y posiblemente los objetivos del análisis.

Anexo

MATRIZ DE OBSERVACIONES

166.02	64.19	0.82	1577.21	56.33	5.28	8.46	2.72	2.68
180.65	99.10	0.80	2041.40	59.00	4.74	5.12	3.08	0.80
170.76	63.09	0.89	1423.00	82.63	0.43	2.24	2.15	0.0
80.19	43.83	0.87	1008.14	36.95	10.57	19.81	1.06	6.17
156.17	45.75	0.50	500.26	76.31	13.10	8.53	1.10	2.42
274.08	132.52	0.57	1418.76	65.90	3.36	9.28	1.91	0.0
415.88	195.71	0.62	1330.80	95.74	2.89	1.13	1.64	0.0
623.10	247.64	0.58	2146.22	127.92	2.59	11.82	1.92	3.26
150.60	82.86	0.83	6325.00	58.38	1.98	3.06	1.86	0.0
1418.50	292.04	0.43	1985.90	151.60	0.96	4.97	1.09	0.0
1058.66	349.82	0.40	2117.32	194.34	1.06	2.99	1.32	0.10
1703.53	433.10	0.34	5110.60	250.52	0.28	1.63	1.33	0.0
614.05	242.16	0.56	1121.30	104.84	0.47	11.37	1.59	0.84
215.20	103.14	0.88	726.29	55.51	5.01	3.10	1.52	8.70
102.53	78.26	0.99	952.88	37.50	14.79	1.80	1.39	3.17
237.90	137.09	0.68	1617.70	51.85	12.11	10.20	1.22	5.26
111.29	69.42	0.92	1168.50	51.93	11.87	2.46	1.78	0.0
3540.60	309.67	0.09	3662.69	216.77	2.21	6.81	1.07	1.40
467.06	184.41	0.54	961.06	87.68	1.97	6.73	1.48	4.37
504.51	179.69	0.64	2810.86	121.08	1.05	9.53	1.45	0.08
388.92	185.84	0.73	4191.67	96.48	0.75	4.28	1.79	1.39
2321.31	413.38	0.18	3353.00	212.51	1.00	2.74	1.40	0.0
435.98	246.51	0.75	5861.44	125.90	0.66	0.97	1.59	0.0
527.70	303.68	0.67	4023.75	112.16	0.49	0.83	1.98	0.0
766.94	343.34	0.56	1018.17	171.67	0.28	1.16	1.94	0.0
602.03	252.97	0.64	4087.47	147.37	1.00	2.84	1.73	0.0
389.92	221.09	0.82	4289.10	126.90	0.62	0.58	1.77	0.0
166.95	65.23	0.88	1869.87	58.56	1.65	2.66	2.95	0.0
465.86	185.39	0.78	2595.50	139.22	0.55	1.84	1.88	0.0
321.84	207.34	0.68	2695.38	96.26	0.29	4.96	2.30	0.0
1342.94	253.67	0.21	2853.75	211.39	0.34	7.42	2.03	3.23
269.23	105.89	0.68	1100.35	50.21	3.40	11.05	1.27	4.03
237.81	66.51	0.84	1388.80	51.67	2.55	10.30	1.30	7.05
358.52	174.07	0.68	6005.25	109.68	0.60	1.69	2.16	0.0
186.52	93.92	0.84	1324.30	55.41	1.91	3.75	2.42	6.12
275.68	173.53	0.71	3959.73	85.57	0.94	2.48	2.80	0.0
211.56	76.47	0.91	2307.91	71.11	1.39	4.54	2.76	1.65
277.02	100.32	0.89	2216.18	84.94	0.77	4.54	1.92	0.0
302.83	86.52	0.89	7268.00	88.63	0.58	5.89	2.61	0.0
407.90	128.03	0.89	3897.67	108.60	0.82	15.39	2.71	0.0
150.60	58.57	1.00	2635.50	65.89	1.72	2.45	3.44	0.0
126.75	48.00	0.89	1806.13	50.00	5.16	1.93	3.71	0.0
152.84	57.02	0.85	1069.88	47.24	4.99	9.34	1.11	8.24
72.55	36.17	0.96	798.07	25.36	7.23	5.60	0.90	12.12
62.92	34.48	0.74	646.80	19.37	9.36	5.77	1.12	27.38
151.39	55.15	0.39	643.42	40.42	4.22	10.37	1.15	4.44
471.31	167.58	0.51	837.89	96.68	1.11	6.60	1.22	3.39

MATRIZ NORMALIZADA

−0.53	−0.91	0.59	−0.52	−0.74	0.55	0.69	1.31	0.04
−0.51	−0.58	0.50	−0.24	−0.69	0.40	−0.11	1.84	−0.37
−0.53	−0.93	0.92	−0.61	−0.25	−0.74	−0.81	0.46	−0.54
−0.67	−1.11	0.82	−0.86	−1.10	1.95	3.43	−1.17	0.79
−0.55	−1.09	−0.89	−1.16	−0.37	2.62	0.71	−1.11	−0.02
−0.36	−0.26	−0.56	−0.61	−0.56	0.04	0.89	0.10	−0.54
−0.14	0.35	−0.33	−0.66	−0.01	−0.09	−1.08	−0.30	−0.54
0.19	0.85	−0.52	−0.18	0.58	−0.17	1.50	0.11	0.16
−0.56	−0.74	0.64	2.32	−0.70	−0.33	−0.61	0.02	−0.54
1.45	1.27	−1.21	−0.27	1.02	−0.60	−0.15	−1.12	−0.54
0.88	1.83	−1.35	−0.19	1.81	−0.57	−0.63	−0.78	−0.52
1.90	2.63	−1.63	1.60	2.84	−0.78	−0.96	−0.77	−0.54
0.18	0.79	−0.61	−0.79	0.16	−0.73	1.39	−0.38	−0.36
−0.46	−0.54	0.87	−1.03	−0.75	0.48	−0.60	−0.48	1.33
−0.63	−0.78	1.38	−0.89	−1.09	3.07	−0.92	−0.68	0.14
−0.42	−0.21	−0.06	−0.49	−0.82	2.36	1.11	−0.93	0.59
−0.62	−0.86	1.06	−0.76	−0.82	2.29	−0.76	−0.09	−0.54
4.81	1.44	−2.79	0.73	2.22	−0.27	0.29	−1.15	−0.24
−0.06	0.24	−0.70	−0.89	−0.16	−0.33	0.27	−0.54	0.40
0.00	0.19	−0.24	0.22	0.45	−0.57	0.95	−0.59	−0.52
−0.18	0.25	0.18	1.05	0.00	−0.65	−0.32	−0.08	−0.24
2.88	2.44	−2.37	0.55	2.14	−0.59	−0.69	−0.66	−0.54
−0.11	0.84	0.27	2.05	0.54	−0.68	−1.12	−0.38	−0.54
0.04	1.38	−0.10	0.95	0.29	−0.72	−1.15	0.20	−0.54
0.42	1.77	−0.61	−0.85	1.39	−0.78	−1.07	0.14	−0.54
0.16	0.90	−0.24	0.99	0.94	−0.59	−0.66	−0.17	−0.54
−0.18	0.59	0.59	1.11	0.56	−0.69	−1.21	−0.11	−0.54
−0.53	−0.90	0.87	−0.34	−0.70	−0.41	−0.71	1.65	−0.54
−0.06	0.25	0.41	0.09	0.79	−0.71	−0.91	0.05	−0.54
−0.29	0.46	−0.06	0.15	−0.00	−0.78	−0.15	0.68	−0.54
1.33	0.90	−2.23	0.25	2.12	−0.76	0.44	0.28	0.15
−0.37	−0.51	−0.06	−0.80	−0.85	0.05	1.32	−0.85	0.33
−0.42	−0.89	0.69	−0.63	−0.83	−0.18	1.13	−0.81	0.98
−0.23	0.14	−0.06	2.13	0.24	−0.69	−0.94	0.47	−0.54
−0.50	−0.63	0.69	−0.67	−0.76	−0.35	−0.45	0.86	0.78
−0.36	0.14	0.08	0.91	−0.20	−0.60	−0.75	1.43	−0.54
−0.46	−0.80	1.01	−0.08	−0.47	−0.48	−0.25	1.37	−0.19
−0.36	−0.57	0.92	−0.13	−0.21	−0.65	−0.25	0.11	−0.54
−0.32	−0.70	0.92	2.89	−0.14	−0.70	0.07	1.14	−0.54
−0.15	−0.30	0.92	0.87	0.22	−0.63	2.36	1.29	−0.54
−0.56	−0.97	1.43	0.12	−0.56	−0.40	−0.76	2.38	−0.54
−0.60	−1.07	0.92	−0.38	−0.86	0.52	−0.88	2.78	−0.54
−0.55	−0.98	0.73	−0.82	−0.91	0.47	0.90	−1.09	1.23
−0.68	−1.18	1.24	−0.98	−1.31	1.06	0.00	−1.41	2.07
−0.70	−1.20	0.22	−1.07	−1.42	1.63	0.04	−1.08	5.35
−0.56	−1.00	−1.40	−1.08	−1.03	0.27	1.15	−1.03	0.41
−0.05	0.08	−0.84	−0.96	0.00	−0.56	0.24	−0.93	0.19

CAPITULO 9

Lista de participantes.

Lista de participantes

9

Ing. Agr. Juan Algorta
Técnico
OPYPA
Montevideo

Ing. Agr. Alfredo Alonso Elizondo
Encargado del Departamento de Estadísticas de la Producción
DIEA
Montevideo

Sr. Guillermo Artigue
Técnico
DIEA
Montevideo

Ing. Agr. Darío Cal
Director
DIEA
Montevideo

Ing. Agr. Roberto Casás Bernadá
Especialista en Proyectos Agrícolas
IICA
Montevideo

Ing. Agr. Miguel Cetrángolo
Economista Agrícola
IICA
Asunción, Paraguay

Ing. Agr. Roberto Jorge Claramunt Sapriza
Técnico
CIAAB – DIEA
Montevideo

Dr. Hugo E. Cohan
Especialista en Economía Agrícola
IICA
Montevideo

Ing. Agr. Humberto Costa Fernández
Técnico
OPYPA
Montevideo

Ing. Agr. Raúl Chiesa
Técnico
CIAAB – DIEA
Montevideo

Ing. Agr. Martín Juan Dabezies Antía
Técnico
CIAAB
Montevideo

Sr. Néstor Eulacio
Experto en Estadística
DIEA
Montevideo

Ing. Agr. Daniel H. Faggi
Jefe del Proyecto Nacional de Lechería
CIAAB
Montevideo

Ing. Agr. José María Ferrari
Encargado de la Sub-Dirección de Estudios Econométricos
DIEA
Montevideo

Prof. Pedro E. Ferreira
CIENES – OEA
Santiago, Chile

Sr. Juan M. Gallo
Computador Universitario
Div. Computación, Univ. de la República
Montevideo

Sr. Daniel E. Gascue
Computador Universitario
Encargado Dirección Unidad de Computación
Min. de Educación y Cultura
Montevideo

Ing. Agr. Tomás Backer Ecos González
Especialista en Economía Agrícola
IICA
Brasil

Sr. Wilfredo A. Ibáñez
Técnico Auxiliar
CIAAB – La Estanzuela
Colonia

Dr. Eduardo Indarte
Jefe Alterno Depto. de Economía
Plan Agropecuario
Montevideo

Ing. Agr. Rodolfo M. Irigoyen
Asesor Técnico
DINACOSE
Montevideo

Dr. Mario Kaminsky
Profesor, Coordinador Prog. Inv. Aplicada
CIENES – OEA
Santiago, Chile

Lic. Beatriz Licio
Becario IICA (FSB)
Montevideo

Ing. Agr. Emilio Montero
Coord. Plan de Acción de IICA en Uruguay
IICA
Montevideo

Sr. Jorge A. Moretti
Encargado (Int.) del Depto. de Censos y Encuestas
DIEA
Montevideo

Ing. Agr. Daniel Nicola
Técnico
Secretariado Uruguayo de la Lana
Montevideo

Ing. Agr. Raúl Oficialdegui
Técnico
Secretariado Uruguayo de la Lana
Montevideo

Ing. Agr. Walker Pascale
Profesor de Adm. Rural - Fac. de Agronomía
Técnico – DIEA
Montevideo

Ing. Agr. Carlos A. Peixoto
Técnico
OPYPA
Montevideo

Ing. Agr. Jorge Pereira Darriufat
Particular

Ing. Agr. Ezequiel Pérez Alvarez
Encargado de Extensión
Secretariado Uruguayo de la Lana
Montevideo

Ing. Agr. Carlos Pérez Arrarte
Asesor Económico
DINACOSE
Montevideo

Cdora. Susana L. Picardo Prats
Dr. Depto. Estadísticas Continuas
Dirección Gral. de Estadística y Censos
Montevideo

Sr. Félix A. Pimentel
Computador Universitario
Banco de la República
Montevideo

Dr. Martín Piñeiro
Coord. Proyecto Cooperativo de Invest. sobre Tecnología
Agropecuaria en América Latina
IICA
Colombia

Ing. Agr. Domingo Quintans
Asistente Cátedra de Adm. Rural
Facultad de Agronomía
Montevideo

Ing. Agr. José María Reyes Amaro
Técnico Adjunto
DIEA
Montevideo

Ing. Agr. Oscar Sarroca
Técnico
CIAAB
Montevideo

Ing. Agr. Jorge E. Schenone
Asistente de Cátedra de Adm. Rural
Facultad de Agronomía
Montevideo

Ing. Agr. Alfredo Weiss
Profesor Titular de Economía Agraria
Facultad de Agronomía
Montevideo

Depósito Legal No. 129.147

Imp. I.G.M.

SERIE DE INFORMES DE CONFERENCIAS, CURSOS Y REUNIONES No. 136