Universidad Nacional Autónoma de México
Facultad de Ingeniería
División de Estudios de Postgrado
Departamento de Sistemas

"Análisis Estadístico Multivariante"

Tesis para la Obtención del Grado de Maestro en Ingeniería

(Investigación de Operaciones)

Ing. Carlos Barcia Portillo

Director de tesis:
M.I. Rubén Téllez Sánchez

Diciembre 200

El análisis de los métodos multivariantes predominará en el futuro y dará por resultado cambios drásticos en el modo en que los investigadores piensan sobre los problemas y como diseñan sus investigaciones.

Hardyck y Petrinovich.

Agradecimientos.

Con este trabajo culmino una etapa muy importante de mi vida y se lo debo a muchas personas, muy especialmente a Fernando Vázquez Gutiérrez y su familia, principal apoyo moral y material para la realización de mis estudios de maestría, e incondicional amigo.

A Raimundo Franco promotor e inspirador para mi superación personal.

Rubén Téllez ha contribuido también con todos sus conocimientos y su invaluable amistad durante toda la maestría y en especial en este trabajo, a el mi infinito agradecimiento.

A las autoridades y el personal de la UNAM, en especial a la Dirección de Apoyo al Posgrado y a la Dirección de Intercambio Académico sin cuyo apoyo no hubieran sido posible mis estudios en México.

Mi agradecimiento también para todos los profesores de la División de Estudios de Posgrado de la Facultad de Ingeniería de la UNAM, por ser fuente inagotable de conocimientos y por ponerlos a mi disposición.

Mercedes Delgado es otra de las personas a quién mucho agradezco por su colaboración a lo largo de tantos años de estudio y trabajo conjunto, siempre dispuesta a aportar sus conocimientos y experiencias de forma desinteresada.

A la dirección del Instituto de Cibernética Matemática y Física que propició estos estudios.

Finalmente a mi familia, pues sin su apoyo y comprensión no hubiera podido llegar hasta aquí y tener fuerzas para seguir adelante.

Carlos Barcia Portillo

México, D.F

Contenido.

Introducción.

Uno de los aspectos fundamentales para lograr el éxito en un proceso de toma de decisiones o en una investigación científica, incluyendo las relacionadas con la investigación de operaciones, es la posibilidad de información y el análisis adecuado de la misma. En muchas ocasiones es necesario decidir sobre el lanzamiento de un nuevo producto, la adopción de una determinada política económica, la adquisición de un inmueble, etc. En estos casos y en muchos otros es muy importante disponer de la información básica (datos) y considerar técnicas estadísticas que faciliten el análisis de la misma.

Los datos que por lo general obtenemos sobre el fenómeno objeto de estudio manifiestan un conjunto de factores que interactúan simultáneamente sobre él, por lo que es necesario considerar métodos de estudio que traten el análisis múltiple de variables o análisis multivariante.

En los últimos años las Técnicas de Análisis Multivariante de datos han tenido una amplia aceptación y uso en la industria, la administración, los negocios y en casi todos los campos de la investigación científica debido, entre otras, a las siguientes razones.

- La comprobación de que en muchas investigaciones y análisis es necesario considerar las relaciones simultaneas entre tres o más variables.

- Los grandes progresos de las computadoras, con gran capacidad de almacenamiento y de procesamiento de información, y el desarrollo de software de gran eficiencia y fácil uso para la implementación del análisis multivariante.

Si bien se ha dado un auge en estas técnicas es necesario señalar que su principal desarrollo, en cuanto a publicaciones, se ha alcanzado en el campo teórico – matemático y no en el destinado a las aplicaciones.

Conforme a lo anterior el principal objetivo de este trabajo puede definirse de la forma siguiente.

> *Abordar un conjunto de las principales técnicas de análisis multivariante disponibles, con un enfoque en los aspectos metodológicos y conceptuales, que permitan su mejor comprensión y facilite su aplicación por parte de profesionales de diferentes campos, sin necesidad de ser especialistas en estadística o en matemáticas.*

Por otro lado, como se ha planteado anteriormente, uno de los factores vitales en el desarrollo de las técnicas de análisis multivariante ha sido el desarrollo de la informática y la computación, que ha puesto a disposición de todos numerosos paquetes estadísticos que permiten la solución de cualquier problema multivariante.

Con vistas a que esta facilidad pueda ser mejor aprovechada, nos proponemos como un objetivo complementario en nuestra tesis el siguiente.

> *Realizar una relación de los principales paquetes estadísticos que tratan el análisis multivariante y brindar un análisis de los principales opciones que con este propósito brinda el paquete SPSS . Ejemplificando su utilización en casos de aplicación.*

Este trabajo comienza necesariamente definiendo en su primer capítulo que entendemos por análisis multivariante. En este capítulo se recoge además un resumen de las principales técnicas de análisis multivariante que serán tratadas a lo largo de la tesis y se propone una metodología para su aplicación.

Teniendo en cuenta las etapas que establece dicha metodología, en el capítulo 2 se discuten los aspectos relacionados con el análisis previo de los datos; en el capítulo 3, se establece una clasificación de las técnicas multivariante bajo estudio y se tratan los supuestos básicos para su aplicación.
El capítulo 4 aborda cada técnica con mayor profundidad, buscando reflejar los aspectos básicos para su aplicación y evitando los aspectos más complejos de notaciones estadísticas y matemática, aunque conservando la rigurosidad teórica.

El capítulo 5 se refiere a los principales paquetes estadísticos que abordan el análisis multivariante, en particular el paquete SPSS, el cual es utilizado en la solución de tres problemas concretos en el capítulo 6

Finalmente se proponen las conclusiones y recomendaciones del trabajo.

Capítulo 1

Análisis multivariante de datos.

l capítulo 1 constituye una introducción al tema del análisis multivariante y un punto de partida para la comprensión de los principales aspectos tratados en el presente trabajo. En el se define que se entiende por análisis multivariante de datos, se establece una revisión simplificada de las principales técnicas a tratar y por último se propone una metodología general para su aplicación.

1.1. Análisis multivariante de datos. Definición.

El término multivariante no es utilizado de la misma forma en la literatura. Para algunos autores multivariante significa simplemente examinar relaciones entre más de dos variables. Otros usan el término sólo para problemas en los que se supone que todos las variables tienen una distribución normal multivariante. Sin embargo para ser considerado verdaderamente multivariante, todas las variables deben ser aleatorias y estar relacionadas de tal forma que sus diferentes efectos no puedan ser interpretados separadamente con algún sentido. Algunos autores afirman que el propósito del análisis multivariante es medir, explicar y predecir el grado de relación de las combinaciones ponderadas de variables. Por lo tanto, el carácter multivariante reside en las combinaciones múltiples de variables y no sólo en el número de variables u observaciones.

A nuestro efecto, el análisis multivariante incluirá tanto técnicas multivariable como multivariante, debido a que el conocimiento de las técnicas multivariable es un primer paso esencial en la comprensión del análisis multivariante.

A partir de las consideraciones anteriores podemos definir el ***análisis multivariante*** en un sentido amplio como el "***conjunto de métodos estadísticos que analizan simultáneamente medidas múltiples de cada individuo u objeto sometido a investigación" [Hair, 1999]***, o en otros términos, como ***"la aplicación de métodos que tratan con un número razonablemente grande de mediciones (variables) hechas en cada objeto, en una o más muestras simultáneamente" [Dillon, 1984],*** de donde indudablemente se desprende como punto importante que el análisis multivariante tiene que ver con las relaciones simultaneas entre variables. Aunque cualquier análisis simultaneo de más de dos variables puede ser considerado aproximadamente como un análisis multivariante.

1.2. Principales técnicas de análisis multivariante.

La variedad de técnicas de análisis multivariante existentes es muy amplia y sigue en expansión. Trataremos un grupo de ellas que consideramos entre las más conocidas y aplicadas, de las cuales se dará a continuación brevemente su definición y objetivo, y en el Capítulo 3 se tratarán con mayor profundidad.

Técnicas Factoriales: Estos métodos consisten en condensar la información contenida en un número de variables originales en un número menor de nuevas variables creadas por el propio análisis, que contienen sin embargo gran parte de la información. Puede usarse para analizar interrelaciones entre un gran número de variables y explicar estas variables en términos de sus dimensiones (factores). Incluye técnicas como el análisis de componentes principales, el análisis factorial y el análisis de correspondencia.

- **Análisis de Componentes Principales (ACP):** Reduce las variables originales a unas nuevas variables sintéticas llamadas componentes o factores que se caracterizan por estar incorrelacionadas entre sí. Cada una de estas componentes se ordenan de acuerdo con la información que contienen, cuantificadas a través de su varianza.

- **Análisis Factorial:** El planteamiento de este método parte de una matriz de correlaciones entre p variables y de la hipótesis de que dichas correlaciones no son aleatorias, sino que se debe a que las variables comparten causas comunes, llamadas factores. El objetivo del análisis factorial es precisamente identificar dichos factores comunes y cuantificarlos.

- **Análisis de Correspondencia:** A diferencia de los dos anteriores es un método multivariante de reducción de la dimensión válido para variables cualitativas. Esta técnica es muy usada en trabajos de investigación de mercados, especialmente en estudios de imagen de marcas y posicionamiento de productos.

Análisis de Grupos (Cluster) : Es un método multivariante de clasificación de datos, cuyo objetivo es formar grupos de individuos homogéneos y mutuamente excluyentes respecto a un conjunto de características que pueden ser cualitativas o cuantitativas, estos grupos no están predefinidos por lo que se usa la técnica para identificarlos. Existen diversos algoritmos cluster y dependiendo del tipo de datos manejados y de los objetivos concretos del estudio se aplica uno u otro.
Una aplicación muy extendida de esta técnica es la segmentación de mercados, consistente en la agregación de consumidores en grupos homogéneos. Un ejemplo sería la determinación de turistas tipo.

Análisis Multidimensional: El objetivo del análisis multidimensional es transformar las opiniones de preferencia de los consumidores en distancias representadas en un espacio multidimensional. Si los objetos A y B son en opinión de los encuestados más similares que el resto de los pares posibles de objetos, esta técnica los situará de tal forma que la distancia entre ellos en un espacio multidimensional es menor que la distancia entre cualesquiera otros dos puntos. Además de mostrar el posicionamiento relativo entre los objetos, se realiza un análisis adicional para evaluar que atributos predicen la posición de cada objeto

Análisis de Regresión Múltiple: Es un método de análisis apropiado para cuando el problema del investigador incluye una única variable cuantitativa dependiente que se supone está relacionada con una o más variables cuantitativas independientes. Su objetivo es predecir los cambios en la variable dependiente en respuesta a cambios en varias de las variables independientes. Un ejemplo de aplicación podría ser la predicción de las ventas de

una compañía a partir de información sobre sus gastos de publicidad, número de vendedores y número de distribuidores de su producto.

Análisis Discriminante: Tiene como objetivo explicar la pertenencia de los individuos a grupos previamente establecidos, así como asignar nuevos individuos a dichos grupos en función de las variables utilizadas para definir un criterio de clasificación. Una aplicación típica de esta técnica puede ser el criterio de concesión de créditos bancarios a los clientes. Teniendo en cuenta el perfil socio-económico y demográfico del solicitante y su situación financiera se decide la posible concesión del crédito según su pertenencia al grupo de demandantes que amortizará el crédito o al de los que no lo amortizarán.

Análisis Multivariante de la Varianza: El análisis multivariante de la varianza (MANOVA) puede ser usado para explorar simultáneamente relaciones entre diversas categorías de variables independientes y dos o más variables métricas dependientes.
MANOVA es una técnica de dependencia que mide la diferencia de dos o más variables métricas dependientes basadas en un conjuntos de variables no métricas que actúan como predictoras. Se emplea para contrastar la significación estadística de las diferencias entre los grupos.
Esta técnica es particularmente útil cuando el investigador diseña una situación experimental para comprobar hipótesis concernientes a la varianza de respuestas de grupos sobre dos o más variables métricas.

Análisis de Correlación Canónica: Esta técnica puede considerarse una extensión de un análisis de regresión múltiple. Su objetivo es correlacionar simultáneamente varias variables dependientes métricas y varias variables métricas independientes. Se busca desarrollar una combinación lineal de cada conjunto de variables (tanto dependientes como independientes) para maximizar la correlación entre los dos conjuntos.

Análisis Conjunto: Es una técnica de dependencia emergente que ha introducido una nueva sofisticación en la evaluación de objetos, sean nuevos productos, servicios o ideas. La aplicación más directa está en productos nuevos o desarrollo de servicios, permitiendo la evaluación de productos complejos mientras que mantiene un contexto de decisión realista para el encuestado.

Análisis Logit: Esta técnica consiste en una combinación de la regresión múltiple y el análisis discriminante múltiple, ya que en ella se usan una o más variables independientes para predecir una única variable dependiente, como en el análisis de regresión múltiple, pero con la particularidad de que la variable dependiente es cualitativa, como en el análisis discriminante. Distinguiéndose de este último en que acomoda todos los tipos de variables independientes (cualitativas y cuantitativas) y no requiere el supuesto de normalidad multivariante.

Modelo de Ecuaciones Estructurales (SEM): Esta técnica constituye una extensión de varias técnicas de análisis multivariante, entre ellas la regresión múltiple y el análisis factorial. El SEM examina simultáneamente una serie de relaciones de dependencia. Este conjunto de relaciones cada una con variables dependientes e independientes es la base del SEM. Abarca una familia de modelos conocidos con muchos nombres, entre ellos análisis

de la estructura de la covarianza, análisis de variable latente, análisis de factor confirmatorio y a menudo simplemente LISREL (el nombre de uno de los paquetes de software más populares).
En la tabla 1.1 se relacionan a modo de ejemplo algunas de las aplicaciones que en diversos campos han tenido estas técnicas.

Tabla 1.1. Ejemplos de aplicación de las diversas técnicas abordadas.

ANÁLISIS DE COMPONENTES PRINCIPALES ANÁLISIS FACTORIAL	Resumir información que aparece en un amplio cuestionario, reduciendo los problemas que se asocian con las grandes cantidades de variables o altas intercorrelaciones.
ANÁLISIS DE CORRESPONDENCIAS	Comparación directa de similitud y diferencia de empresas y de los atributos asociados. Estudio de imagen de marcas y posicionamiento.
ANÁLISIS CLUSTER	Clasificación de datos. Ejemplo: Segmentación de mercados. Consistente en la agregación de consumidores en grupos homogéneos como puede ser la determinación de turistas tipo
ANÁLISIS MULTIDIMENSIONAL	En marketing es empleada para identificar los criterios determinantes en las evaluaciones de los servicios o compañías por parte de los clientes. Identificar segmentos de mercados basándose en juicios de preferencia. Determinar que productos son más competitivos entre sí (es decir son más parecidos) Deducir que criterios usa la gente cuando juzga los objetos.. Evaluación de diferencias culturales entre diferentes grupos.
ANÁLISIS DE REGRESIÓN MÚLTIPLE	Aplicable en cualquier ámbito de la toma de decisiones en los negocios. Son la base de los modelos de previsión económica. Ejemplos: modelos de comportamiento de una empresa en el mercado si se sigue una estrategia de marketing determinada, predicción de ventas de una compañía a partir de la información de sus gastos en publicidad, número de vendedores y número de tiendas que distribuyen sus productos, etc.
ANÁLISIS DISCRIMINANTE MÚLTIPLE	Determinar pertenencia a un grupo. Ejemplos: Para un cliente: Criterio de concesión de créditos bancarios a los clientes(credit scoring bancario) Diagnóstico de enfermedades. Para un producto: Éxitos en ventas frente a fracasos. Distinción entre usuarios habituales u ocasionales de un producto Para una empresa: Rentable frente a no rentable.

ANÁLISIS MULTIVARIANTE DE LA VARIANZA	Estudio de las influencias de los métodos de docencia en los resultados académicos de los estudiantes
ANÁLISIS DE CORRELACIÓN CANÓNICA	Tratamiento de múltiples variables dependientes. Ejemplo: Predecir el uso del crédito utilizando como variables dependientes el número de tarjetas de crédito que tienen las familias y el gasto promedio mensual con las tarjeta y como variables independientes el tamaño y la renta d la familia. Determinación de la relación entre las percepciones de un cliente sobre una empresa y el nivel de uso de satisfacción del cliente
ANÁLISIS CONJUNTO	Evaluación de nuevos productos, servicios o ideas
ANÁLISI LOGIT	Identificación del grupo al cual un objeto(persona, empresa, producto, etc) pertenece. Ejemplos: predicción de éxito o fracaso de un nuevo producto, clasificación de estudiantes según sus intereses vocacionales, predicción de éxito para una empresa
MODELO DE ECUACIONES ESTRUCTURALES	Utilizado en casi todos los campos de estudio, incluyendo la educación, el marketing, la psicología, la sociología, la salud, la demografía, el comportamiento organizacional, la biología y la genética. Una aplicación muy extendida es en el análisis de relaciones dentro de cualquier sistema. Ejemplo: el análisis de las relaciones entre las actitudes de los empleados en una compañía.

1.4. Metodología para la aplicación del análisis multivariante.

El objetivo de muchas de las técnicas a analizar es la descripción simplificada de la estructura de las observaciones o diseño de modelos, lo cual debido a las numerosas técnicas de análisis multivariante disponibles y la gran cantidad de supuestos que implica su aplicación puede resultar una tarea difícil. Una metodología que puede ayudar en el proceso de aplicación de las técnicas multivariante se describe en la figura 1.1

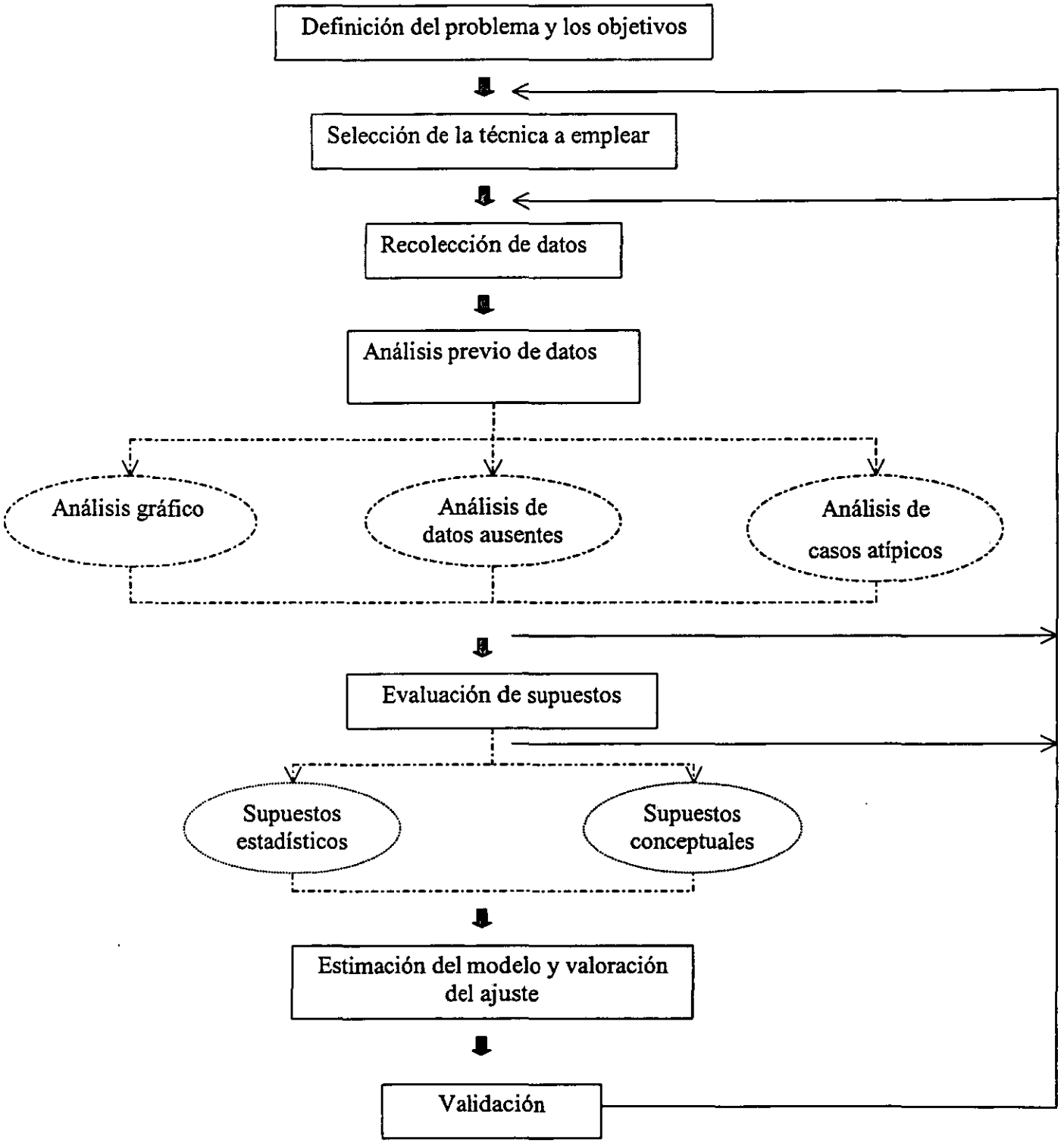


Figura 1.1. Metodología general para la aplicación del análisis multivariante

Esta metodología constituye una propuesta de la forma de abordar el análisis multivariante para facilitar el trabajo de investigadores y profesionales que requieren aplicar estas técnicas. Cada una de las etapas establecidas serán abordadas a continuación brevemente y aquellas que de acuerdo a los propósitos de nuestro trabajo lo requieran serán tratadas con mayor detalle en los capítulos posteriores.

Metodología general para la aplicación del análisis multivariante

Definición del problema y los objetivos.
El primer paso en cualquier análisis multivariante debe ser la definición conceptual del problema a investigar y de los objetivos a alcanzar, identificando las relaciones fundamentales a estudiar ya sean relaciones de dependencia o de interdependencia. Lo esencial en esta etapa es identificar las ideas o temas de interés, con lo cual se minimice la posibilidad de que conceptos relevantes sean omitidos en los esfuerzos por desarrollar medidas y definir los detalles del modelo.
Los objetivos que de forma más general pueden satisfacerse con la aplicación de estas técnicas pueden ser recogidos en alguno de los tres grupos siguientes.

1. Reducción de dimensiones (examinar las pautas subyacentes o las relaciones para un amplio numero de variables y determinar si la información puede ser condensada o resumida)
2. Desarrollo y estudio de dependencia (un análisis de dependencia es aquel en el que una variable o conjunto de variables es identificado como dependiente y va a ser explicado por otras variables conocidas como variables independientes)
3. Clasificación multidimensional (identificar la estructura entre una serie de variables, observaciones u objetos definidos para clasificarlos en un esquema mas simple, de tal forma que se puedan caracterizar los objetos dentro de los grupos como un total en vez de tener que tratar con cada objeto individualmente.

Selección de la técnica de análisis multivariante a emplear.
Una vez identificado el problema y los objetivos podemos definir la técnica de análisis multivariante adecuada para resolverlos. Después de definir si la técnica a emplear es de dependencia o interdependencia, la decisión final estará en función del tipo y la cantidad de variables. Una vez determinada la técnica adecuada, es necesario considerar algunos aspectos que dependen de cada técnica tales como el tamaño de muestra y los tipos de variables requeridas, entre otros.

Recolección de datos.
Una vez que seleccionamos el diseño de investigación apropiado y la muestra adecuada de acuerdo con nuestro problema de estudio e hipótesis, la siguiente etapa en recolectar los datos sobre las variables involucradas en la investigación. Recolección de datos relevantes a través de observación pasiva o de la experimentación y de acuerdo a las características requeridas por el modelo.

Recolectar los datos implica tres actividades estrechamente vinculadas entre sí.

1. Seleccionar un instrumento de medición o de recolección de los datos o desarrollar uno. (escala, encuesta, etc)
2. Aplicar ese instrumento y obtener las observaciones y mediciones de las variables de interés para el estudio (medir variables)
3. Preparar las mediciones obtenidas para que puedan analizarse correctamente. (codificación de los datos)

Análisis previo de los datos.
El análisis previo de los datos constituye una parte complementaria del análisis multivariante, que permite una comprensión básica de los datos y las relaciones entre variables.
El conocimiento de las interrelaciones entre variables puede ayudar enormemente en la especificación y refinamiento del modelo multivariante a utilizar, así como proporciona una perspectiva razonable para la interpretación de los resultados. Por otro lado el efecto de los datos ausentes y la presencia de observaciones atípicas pueden ser determinante, por el impacto que tiene sobre la naturaleza y carácter de los resultados.

Evaluación de los supuestos básicos de la técnica seleccionada.
Todas las técnicas tienen supuestos estadísticos y conceptuales que definen su capacidad para representar relaciones multivariantes. Para las técnicas basadas en la inferencia estadística se deben tener en cuenta los supuestos de normalidad multivariante, linealidad, independencia de los términos de error e igualdad de las varianzas en una relación de dependencia.
Antes de pasar a la estimación del modelo deberá asegurarse que todos los supuestos de la técnica a aplicar se encuentren cumplidos, de lo contrario puede que sea necesario la recolección de nuevos datos o incluso la selección de otra técnica para lograr los objetivos trazados

Estimación del modelo y valoración del ajuste.
En este paso se realiza la estimación efectiva del modelo y después se evalúa el ajuste para averiguar si consigue niveles aceptables sobre los criterios estadísticos, si identifica las relaciones propuestas y si consigue la significación práctica. Aquí debe también determinarse si los resultados están excesivamente afectados por un único o pequeño grupo de observaciones que indican que los resultados pueden ser inestables, esto último buscando garantizar la robustez y estabilidad para poder aplicar los resultados razonablemente a todas las observaciones de la muestra. Este paso puede llevar a la re-especificación de variables y/o a la reformulación del modelo.

Validación del modelo.
Cualquiera que sea la técnica de análisis multivariante empleada debe asegurarse que el modelo estimado sea representativo de la población, este es el objetivo de la validación, que asegura el grado de generalidad de los resultados.

Después de ejecutados todos los pasos y de obtener resultados satisfactorios estaremos en condiciones de emplearlos en análisis e investigaciones futuras.

Capítulo 2

Análisis previo de los datos.

El análisis previo de los datos constituye una parte complementaria del análisis multivariante. Un análisis cuidadoso de los datos conduce a una mejor predicción y una evaluación más precisa de la dimensionalidad.
El propósito de este capítulo es dar una visión general de las técnicas de examen de los datos, que van desde el simple proceso de inspección visual de los gráficos hasta el proceso estadístico multivariante que requiere el análisis de datos ausentes y la detección de casos atípicos.

2.1. Tipos de datos y escalas de medida.

Los datos pueden ser considerados unos de los fluidos vitales de la civilización moderna. Ellos son utilizados para tomar decisiones, para soportar decisiones que ya han sido tomadas, para proporcionar un fundamento de porqué un evento ocurrirá, para hacer predicciones sobre la ocurrencia de eventos y para crear y/o probar modelos de un problema particular.

Los datos pueden definirse como hechos o conceptos conocidos o supuestos y disponibles [Taylor, 1990] y aparecen por todas partes en nuestras vidas cotidianas, muchos atributos son medidos u observados en un conjunto de individuos o sujetos y cada atributo en particular se puede considerar como una variable.
Uno de los principales factores que nos permiten distinguir entre las diferentes técnicas de análisis multivariante, cual será la que debemos utilizar es precisamente el tipo de datos disponible o representados

Tipos de datos.
Los datos estadísticos con que generalmente trabajamos consisten en una serie de mediciones u observaciones hechas a un número de individuos, objetos u otra entidad de interés. Si se hace una sola medición a cada individuo, entonces los datos son llamados univariables, si por el contrario se realiza más de una observación o medición a cada individuo, entonces los datos son llamados multivariables [Kraznowski, 1994].
Existen dos tipos de datos claramente definidos y conocidos como datos cuantitativos (métricos, numéricos) y datos cualitativos (no métricos, categóricos), que a los efectos de nuestro trabajo constituye una clasificación básica.

Datos cualitativos: Son atributos, características o propiedades categóricas que identifican o describen a un sujeto. Describen diferencias en tipo o clase, indicando la presencia o ausencia de una característica o propiedad. Muchas propiedades son discretas porque tienen una característica peculiar que excluye todas las demás. Por ejemplo si un sujeto es hombre, no puede ser mujer. No hay cantidad de género, sólo la condición de ser hombre o mujer.

Un caso particular de datos cualitativos es el que sólo tiene dos categorías posibles y son los llamados binarios.

Datos cuantitativos: Este tipo de datos están constituidos de forma tal que los sujetos pueden ser identificados por diferencias entre grado o cantidad. Reflejan cantidades relativas o grados. Son los más apropiados para casos que involucran cantidades o magnitudes tales como el nivel de satisfacción o la demanda de trabajo.
Se asigna un único valor numérico a cada característica del individuo observado.

Los datos son representados generalmente como una matriz de la forma siguiente.

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Donde x_{ij} representa el valor de la j-ésima variable para el i-ésimo individuo. El número de individuos bajo estudio es representado por n y el número de mediciones tomadas a cada individuo se representa por p.

Tabla 2.1. Ejemplo de matriz de datos para un caso hipotético

	Variables medidas					
Individuo	**X_1**	**X_2**	**X_3**	**X_4**	**X_5**	**X_6**
1	M	60	160	Si	Si	Ninguna
2	M	43	N.C	Si	No	Poca
3	M	25	139	No	Si	Poca
4	M	54	170	No	No	Ninguna
5	M	16	120	No	No	Suficiente
6	F	80	N.C	Si	Si	Ninguna
7	F	N.C	84	No	Si	Poca
8	F	49	110	Si	No	Suficiente
9	F	18	90	Si	No	Suficiente
10	F	22	130	No	No	Poca

X_1: Sexo, variable cualitativa , escala de medida nominal.
X_2: Edad, variable cuantitativa, escala de medida razón.
X_3: Peso, variable cuantitativa, escala de medida razón.
X_4: Adicción al tabaco o la bebida, variable cualitativa, escala de medida nominal.
X_5: Padecimiento de hipertensión, variable cualitativa, escala de medida nominal.
X_6: Practica de ejercicios físicos, variable cualitativa, escala de medida ordinal.

N.C: Dato no conocido.

Una tabulación alternativa es la tabla de contingencia. Utilizada cuando todas las variables son categóricas, da la incidencia en el conjunto de datos para todas las combinaciones de categorías de cada variable.

Escalas de medida.

Las mediciones implican diversos tipos de escalas y la aplicación de muchas de las técnicas multivariante que se analizarán dependen directamente del tipo de medición asumido. Por el tipo de escala es posible determinar la correspondencia entre el valor asignado y la propiedad de los objetos bajo estudio y determinar las posibilidades de realizar operaciones matemáticas con estos números.

Por lo general las variables medidas son de diferentes tipos en correspondencia con alguna de las siguientes escalas de medición.

Tabla 2.2. Escalas de medidas y sus características.

Escala de Medida	Características
Nominal	Los datos se describen en términos de clases. Los números asignados indican pertenencia a uno de los conjuntos mutuamente excluyentes y exhaustivos de clase, lo que no implica un orden Variables categóricas
Ordinal	Los datos permiten definir que objeto tiene más, menos o igual cantidad de un atributo en comparación con otro objeto. Es posible establecer un orden sin que este implique distancia entre los distintos puntos en la escala. Variables categóricas
Intervalo	Permite establecer en que medida o cuanto más un objeto tiene determinado atributo en referencia a otro. Existe igual diferencia entre puntos sucesivos en la escala pero el cero es arbitrario. Variables cuantitativas.
Razón	Nos permite definir un origen que representa la cantidad cero del atributo en cuestión. Las diferencias en valores pueden compararse por su magnitud relativa. Variables cuantitativas.

2.2. Análisis previo de los datos.

A continuación se revisan algunos de los métodos gráficos básicos que ayudan a comprender las características de los datos, particularmente en el sentido multivariante. Dado que en la actualidad la mayoría de los paquetes de programas estadísticos tienen módulos para el desarrollo de las técnicas gráficas de análisis de datos, sólo se expondrán los lo esencial de cada método para que puedan ser entendidos.

2.2.1. Gráfico de Dispersión: Se emplea para examinar las relaciones entre dos variables. Existen muchos tipos, pero un formato que se ajusta particularmente a las técnicas multivariante es la matriz del gráfico de dispersión, en ella se representa el gráfico de dispersión para todas las combinaciones de variables en la porción inferior de la matriz. La diagonal contiene los histogramas de las variables y en la parte superior de la matriz se incluyen las correlaciones correspondientes para que se pueda valorar la correlación representada en cada gráfico. Un ejemplo ilustrativo para este tipo de gráfico se muestra en la siguiente figura.

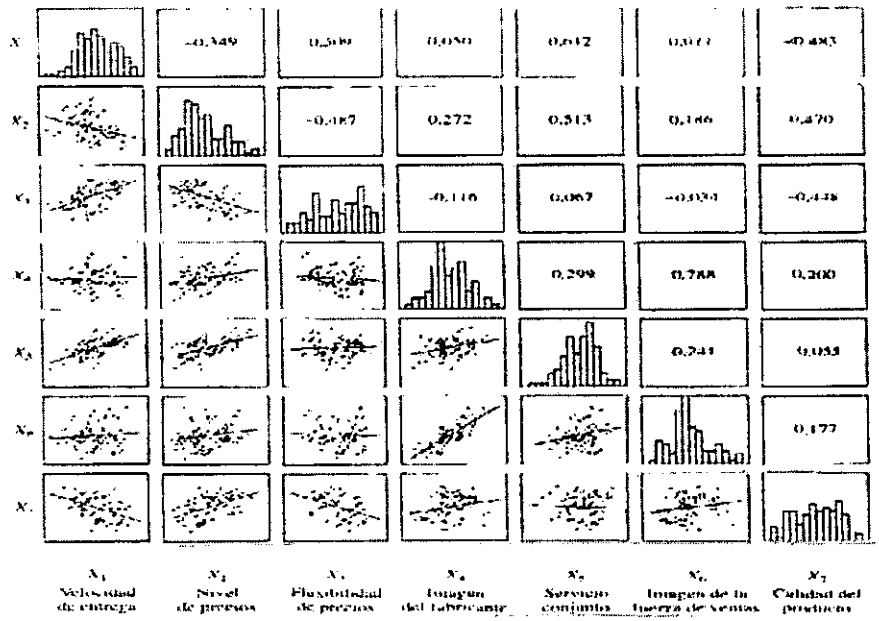


Figura 2.1. Ejemplo de matriz de gráfico de dispersión de variables métricas.

2.2.2. Gráfico de Caja (*Box Plot*): Se emplea para identificar los casos atípicos que pueden detectarse sólo cuando los valores de los datos se separan en grupos. Los límites inferior y superior de la caja marcan los cuartiles inferior y superior de la distribución de los datos. La longitud de la caja es la distancia entre el primer y tercer cuartil, de forma que la caja contiene el 50% de los datos centrales de la distribución. La línea dentro de la caja señala la posición de la mediana. Si esta cae cerca del final de la caja se indica la presencia de asimetría. Cuanto mayor es la caja mayor es la extensión de las observaciones.

Las líneas que se extienden desde cada caja (llamadas bigotes) representan las distancias entre la mayor y menor de las observaciones que están a menos de un cuartil de la caja. Estos valores están marcados con una X. Los casos atípicos son observaciones que se sitúan entre 1.0 y 1.5 cuartiles fuera de la caja.

Los valores extremos son aquellas observaciones mayores que están a 1.5 cuartiles fuera de los límites de la caja.

En la figura 2.2 se puede apreciar un gráfico de caja. Los tres grupos tienen un conjunto de valores muy diferentes, lo que indica que existen verdaderas diferencias entre ellos, en

función de la variable analizada. El gráfico para el primer tipo de situaciones de compra indica además la presencia de un caso atípico.

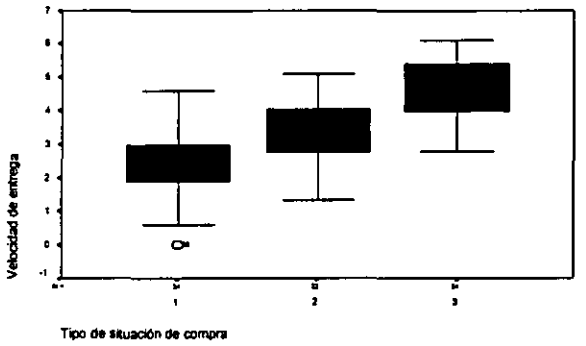


Figura 2.2. Gráfico de caja.

Existen varios métodos gráficos multivariante que permiten la comparación de observaciones caracterizadas por la presencia de más de dos variables, entre las más conocidas tenemos las siguientes.

2.2.3. Gráfico de GLYPH: Ideado por Anderson en 1957 en estos gráficos cada individuo en la muestra es representado por un círculo de radio fijo y cada variable por una recta que parte del circulo. La posición de la recta indica cual variable está siendo representada y la longitud de la recta indica el valor de una variable cuantitativa o la categoría de una variable cualitativa para el individuo representado. Categorías ordenadas (Ej. grande, mediano, pequeño) pueden ser representadas por rectas alargadas progresivamente.
Las categorías más bajas frecuentemente están relacionadas por ausencia de rectas. Si las categorías no están ordenadas el largo de las rectas puede ser asignado arbitrariamente. Para datos cuantitativos el largo de la recta es usualmente proporcional al valor de la variable. El conjunto entero de datos cuantitativos es colocado según una escala donde el valor más pequeño es el cero.

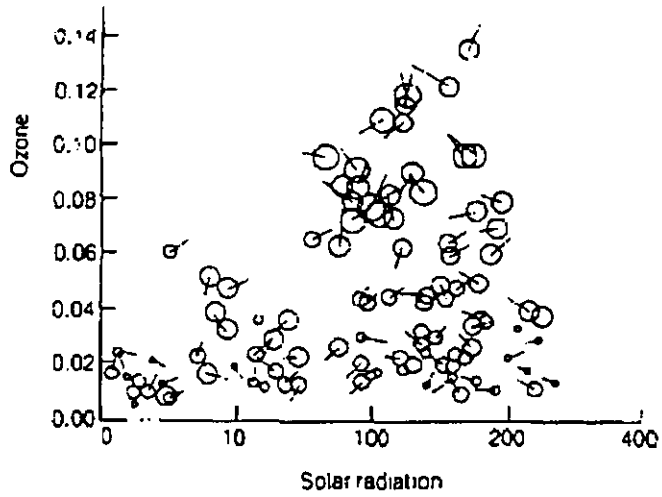


Figura 2.3. Gráfico Glyph. Relación entre el nivel de ozono y la radiación solar.

En la figura 2.3 el gráfico nos brinda información sobre la relación entre el nivel de ozono y la radiación solar , el diámetro de cada círculo se relaciona con la temperatura, la línea que sale del círculo es inversamente proporcional a la velocidad del viento y su orientación indica su dirección. Se aprecia que para un nivel dado de radiación solar, el ozono tiende a incrementarse cuando la temperatura se incrementa y la velocidad del viento disminuye. La dirección del viento no parece tener influencia.

2.2.4. Gráfico de Andrews: El procedimiento es esencialmente muy simple. Cada observación p-dimensional $x' = [x_1, x_2, \cdots, x_p]$ es representada por la curva de Fourier.

$$fx(t) = \frac{x_1}{\sqrt{(2)}} + x_2 \operatorname{sen}(t) + x_3 \cos(t) + x_4 \operatorname{sen}(2t) + x_5 \cos(2t) + \cdots$$

Esta función se plotea en un rango de $-\pi \leq t \leq \pi$. Un conjunto de observaciones multivariante aparece como un conjunto de líneas en el gráfico. La utilidad de esta particular representación se apoya primariamente en el hecho de que esta función preserva la distancia Euclideana, en el sentido que las observaciones estrechamente unidas en el espacio original p-dimensional corresponderán a líneas en el gráfico que permanecen apartadas, al menos para algunos valores de t. Esta propiedad permite a los gráficos ser examinados para distintos grupos de observaciones y observaciones distantes.
Examinando la forma de la función relacionada con el gráfico de Andrews queda claro que las variables originales no son igualmente consideradas. Algunas son asociadas con una componente cíclica teniendo una alta frecuencia, otras con componentes que tienen baja frecuencia.

Andrews en 1972 estableció las siguientes propiedades para estas curvas.

1. La función representada preserva el valor medio, es decir, si $\bar{x}$ es la media de un conjunto de n observaciones multivariantes x_i, entonces en cada punto t en $-\pi \leq t \leq \pi$, la función correspondiente a $\bar{x}$ es la media de las n funciones correspondientes a las observaciones individuales.
2. La función representada mantiene las distancias Euclideanas.
3. Si las variables en la matriz de datos son incorrelacionadas con varianza constante σ^2, entonces las varianzas de la función en cualquier punto t es $\frac{1}{2}\, p\sigma^2$ y $\frac{1}{2}\,(p+1)\sigma^2$ cuando p es par. En cualquiera de los dos casos la dependencia relativa de la varianza de la función en t es muy leve o no existe por lo que la variabilidad de la función graficada es también constante a través del tiempo.
4. La función preserva la relación lineal. Si y está situada en la línea junto a x y z entonces $fy(t)$ aparece junto a $fx(t)$ y $fz(t)$ para todo t.
5. Para un valor particular t_o de t, la función evaluada en t_o es proporcional a lo largo de la proyección del vector x en el vector $a_o = (\frac{1}{2}\sqrt{2}, \operatorname{sen} t_o, \cos t_o, \operatorname{sen} 2t_o, \cos 2t_o, \cdots)$ desde $fx(t_o) = x'a_o$

Dado que en este gráfico los componentes de baja frecuencia son más informativos que los de alta frecuencia, es de mayor utilidad asociar x_1 con la variable considerada en algún sentido la más importante, x_2 con la segunda más importante y así sucesivamente. La detección mediante este método de grupos bien separados podría sugerir la aplicación de algún método de análisis cluster para confirmar la presencia de distintos grupos en el conjunto de datos.

En el gráfico de Andrews representado en la figura 2.4 [Everitt y Dunn, 1991] muestra un ejemplo donde podemos apreciar claramente la presencia de tres grupos bien separados de observaciones, lo cual nos sugiere la aplicación de algún método de análisis cluster

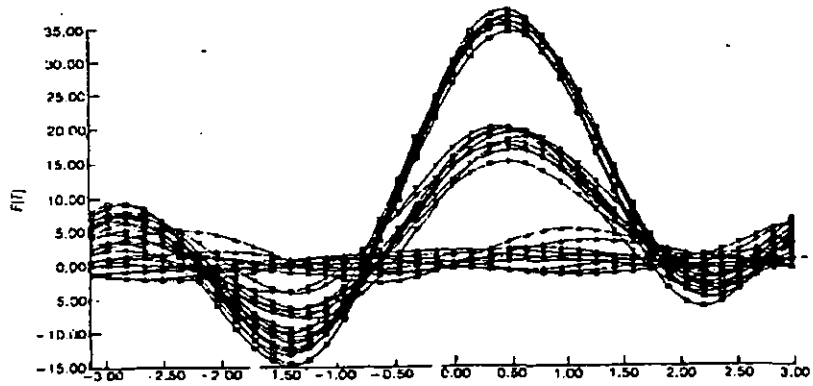


Figura 2.4. Ejemplo de Gráfica de Andrews.

2.2.5. Caras de Chernoff: Chernoff (1973) describe un método para representar datos multivariante en el cual cada observación tiene una cara correspondiente. Las facciones están determinadas por los valores que toma cada variable en particular. Una muestra de observaciones multivariante es representada por una colección de tales caras y estas pueden ser examinadas para detectar similitudes o diferencias entre las observaciones.
Esta técnica ha sido criticada por su subjetividad ya que dado diferentes observaciones es posible establecer diferentes formas para las facciones faciales.
El valor de este tipo de representaciones es la inherente capacidad que tienen los seres humanos para procesar su interpretación.
En el ejemplo de la figura 2.5 se muestra la representación facial de los datos de salarios de la tabla 2.1. En la representación se hace obvia la satisfacción de los trabajadores de Ginebra y Zurcí expresada por su amplia y sonriente cara.
La correspondencia entre las facciones faciales y cada variable es la siguiente:

Maestro	Área de la cara
Chofer	Forma de la cara
Mecánico	Longitud de la nariz
Cocinero	Localización de la boca
Administrador	Curva de la sonrisa
Ingeniero	Ancho de la boca
Cajero	Separación de los ojos

2.3. Valores ausentes en datos multivaridos.

En la práctica es frecuente encontrar en un conjunto de datos multivariante la ausencia de algunos valores. Estas ausencias son perjudiciales no sólo por su potencial sesgo, sino también por su efecto en el tamaño de la muestra disponible para el análisis ya que de no aplicarse alguna solución ninguna observación con datos ausentes sobre cualquier variable puede ser incluida en el análisis.

En tales condiciones se deben buscar observaciones adicionales o encontrar una solución para la ausencia de datos en la muestra original.

Figura 2.5. Representación facial de datos. Caras de Chernoff.
[Everitt, 1991]


Abdhab


Dublin


Luxmb


Sanfm


Amster


Dussel


Madrid


Saoplo


Athens


Geneva


Manila


Singap


Bahran


Helsin


Mexico


Stockh


Bangkk


Hongkn


Milan


Sydney


Bogota


Istanb


Montrl


Tehern


Brussl


Jakart


Newyrk


Telavv


Buenos


Jeddah


Oslo


Tokyo


Caracs


Johann


Panama


Toront


Chicag


London


Paris


Vienna

Copenh

Losang

Riodjn

Zurich

Soluciones para datos ausentes.
Una posible solución en tales casos es excluir del análisis los individuos con valores ausentes. No obstante esta solución debe usarse sólo si los datos ausentes son completamente aleatorios, de lo contrario podrían sesgarse los resultados y no ser generalizables para la población.
Otro problema que puede presentarse al aplicar esta solución es que el tamaño de la muestra resultante quede reducido a una muestra insuficiente para los fines del análisis y la información tendría que ser descartada.

Otra solución simple para los datos ausentes consiste en eliminar las variables que presentan peor comportamiento respecto a los datos ausentes. En este caso debe considerarse lo que se gana al eliminar una fuente de datos ausentes y lo que se pierde al no contar con una determinada variable en el análisis multivariante.

Una solución más satisfactoria podría ser estimar los valores ausentes y entonces completar la matriz de datos con estos valores estimados, un proceso conocido como *imputación*, aplicando entonces la técnica de análisis multivariante con la matriz de datos completada.

Las técnicas de imputación se utilizan fundamentalmente con variables cuantitativas debido a que pueden hacerse estimaciones de los datos ausentes con valores como la media de todos los valores válidos. En el caso de variables cualitativas requieren una estimación de un valor específico en vez de una estimación en una escala continua.

Técnicas de imputación de uso más frecuentes, basadas en la sustitución de los datos ausentes.
Sustitución de casos: En este método las observaciones con datos ausentes se sustituyen con observaciones de otra muestra, preferiblemente muy similar a la de las observaciones originales. Este método es el más utilizado para sustituir las observaciones con datos ausentes completos.

Sustitución por la media: Es uno de los métodos más utilizados. Se basa en el criterio de que la media de los valores válidos de la muestra es el mejor valor de sustitución.

Sustitución por un valor constante: En este caso se sustituye el valor ausente por un valor constante obtenido de fuentes externas o investigaciones previas. Este método puede proporcionar la opción de reemplazar los datos ausentes con valores que podrían ser considerados más válidos que la media de la muestra.

Imputación por regresión: En este método se estiman los valores ausentes mediante un análisis de regresión en el cual, cada variable que contiene valores ausentes es tratada como una variable dependiente y el resto de las variables se consideran variables explicativas, suponiendo que las variables con datos ausentes tienen correlación sustancial con las otras variables.

2.4. Casos atípicos.

Los casos atípicos son observaciones con características que las diferencian claramente de las restantes observaciones. Estas observaciones deben ser consideradas en el contexto del análisis y deben evaluarse por el tipo de información que puede proporcionar para definir finalmente si son beneficiosas o perjudiciales y decidir si las mantenemos o las eliminamos. Si estas observaciones representan a un segmento de la población deberíamos retenerla para asegurarnos de la generalidad de los resultados del análisis.

Causas de ocurrencia de casos atípicos.

1. Observaciones que surgen de un error de procedimiento, tales como la captura de datos o un error de codificación.

2. Observaciones que ocurre como consecuencia de un acontecimiento extraordinario.

3. Observaciones extraordinarias para la cual no se tiene explicación.

4. Observaciones que se sitúan fuera del rango ordinario pero que no son únicas en su combinación de valores entre las variables

Dado que la mayoría de los análisis multivariante tienen más de dos variables, se necesita de una forma de medición objetiva de la posición multidimensional de cada observación con respecto a un punto común. La medida D^2 de Mahalanobis puede usarse con este fin . Se sugiere usar un nivel conservador de 0.001 para designar una observación como caso atípico.

Capítulo 3.

Identificación de las Técnicas de Análisis Multivariante. Evaluación de Supuestos Básicos.

La variedad de métodos multivaridos es muy amplia y la incorrecta selección de la técnica es una causa frecuente de decepciones en su aplicación. Para propiciar una mejor identificación de la técnica que debe ser empleada en cada caso, en este capítulo pretendemos llevar a cabo una clasificación de los métodos que nos ocupan, partiendo de los objetivos que se persiguen con el análisis y de las características de las variables que se utilizan, así mismo, se da una panorámica general de los diferentes supuestos que se deben verifican para una correcta aplicación de estas técnicas.

3.1. Clasificación de las técnicas de análisis multivariante.

Existen una gran variedad de técnicas de análisis multivariante que pueden clasificarse de acuerdo a diferentes aspectos, como son:

a) Si es de interés distinguir entre variables explicativas (independientes) y explicadas (dependientes) o si el interés se centra en la relación mutua entre todas las variables, sin distinción.
b) La naturaleza y número de variables a estudiar.
c) Objetivos del estudio.

El primer aspecto considerado es la división de las variables en explicativas y explicadas, a partir de esta división se determina si se debe emplear una técnica explicativa de dependencia o una técnica descriptiva de interdependencia.

En las técnicas explicativas de dependencia se distinguen entre ambos tipos de variables y se busca explicar o predecir el comportamiento de una o más variables dependientes basados en el conjunto de variables explicativas. En contraste las técnicas descriptivas de interdependencia, que son menos predictivas por naturaleza, no distinguen entre variables explicativas y explicadas, todas las variables son analizadas simultáneamente e intentan encontrar la estructura interna del conjunto de todas las variables para simplificar la complejidad, en primera instancia, a través de la reducción de datos.

Explicación de Dependencia.

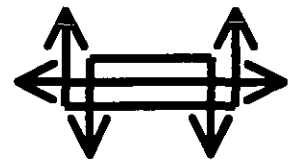
Análisis de Regresión Múltiple.
Análisis Discriminante Múltiple.
Análisis Multivariante de la Varianza
Análisis de Correlación Canónica.
Análisis Conjunto
Modelo de ecuaciones Estructurales

Descripción de Interdependencia.

Técnicas Factoriales.

- Análisis de Componentes Principales
- Análisis Factorial Común
- Análisis de Correspondencias

Análisis de Grupos
Análisis Multidimensional



Las técnicas explicativas de dependencia se pueden, a su vez, dividir según el número de variables dependientes (una variable dependiente, varias variables dependientes o varias relaciones de dependencia – independencia) y de acuerdo al tipo de variable (cualitativa o cuantitativa)

Por otra parte las técnicas descriptivas de interdependencia se subdividen según el objetivo que se persiga (análisis de la estructura de la variable, del objeto o agrupación para representar una estructura) y según el tipo de variable (cualitativa o cuantitativa).

Por último podemos considerar una clasificación más general de las diferentes técnicas multivariante partiendo solamente de los objetivos perseguidos por el investigador, así tenemos:

Objetivo: Reducción de dimensiones
Técnicas: Análisis de Componentes Principales
Análisis Factorial
Análisis de Correspondencia
Análisis Multidimensional

Objetivo: Desarrollo y Estudio de Dependencia Multivariante
Técnicas: Regresión Múltiple
MANOVA
Análisis de Correlación Canónica
Análisis de Ecuaciones Estructurales
Análisis Conjunto
Análisis Discriminante

Objetivo: Clasificación Multidimensional
Técnicas: Análisis Discriminante
Análisis de Grupos
Análisis Multidimensional

En la figura 3.1 se muestra la clasificación de estas técnicas de acuerdo a los criterios expuestos.

3.2. Evaluación de los supuestos básicos de las técnicas de análisis multivariante.

Los procedimientos de análisis multivariante estiman el modelo multivariante aun cuando se incumplen los supuestos estadísticos en lo que se basan, pero estos resultados pueden verse distorsionados y altamente sesgados por esta causa. Debido a esto resulta altamente aconsejable la comprobación del cumplimiento de estos supuestos antes de emprender la aplicación de cualquier técnica de análisis multivariante.

El análisis multivariante requiere que los supuestos estadísticos sean contrastados dos veces, primero para las variables individuales y luego para la combinación ponderada de variables del modelo.

En esta sección nos limitaremos a enunciar los principales test estadísticos aplicables para la comprobación de los supuestos.

Supuesto de Normalidad.
La normalidad de los datos es el supuesto fundamental del análisis multivariante. La normalidad multivariante implica que las variables son normales en un sentido univariante y que sus combinaciones también lo sean.
La normalidad univariante puede diagnosticarse de forma sencilla a través de una comprobación visual del histograma que compara los valores observados con una distribución aproximada a la normal o utilizando test estadísticos que permiten evaluar la normalidad, como es el basado en el valor de simetría Z.
Test estadísticos más específicos se encuentran en la mayoría de los paquetes de programas estadísticos tales como el SPSS y SAS. Los más utilizados son el test de Shapiro-Wilks y una modificación del test de Kolmogorov-Smirnov.

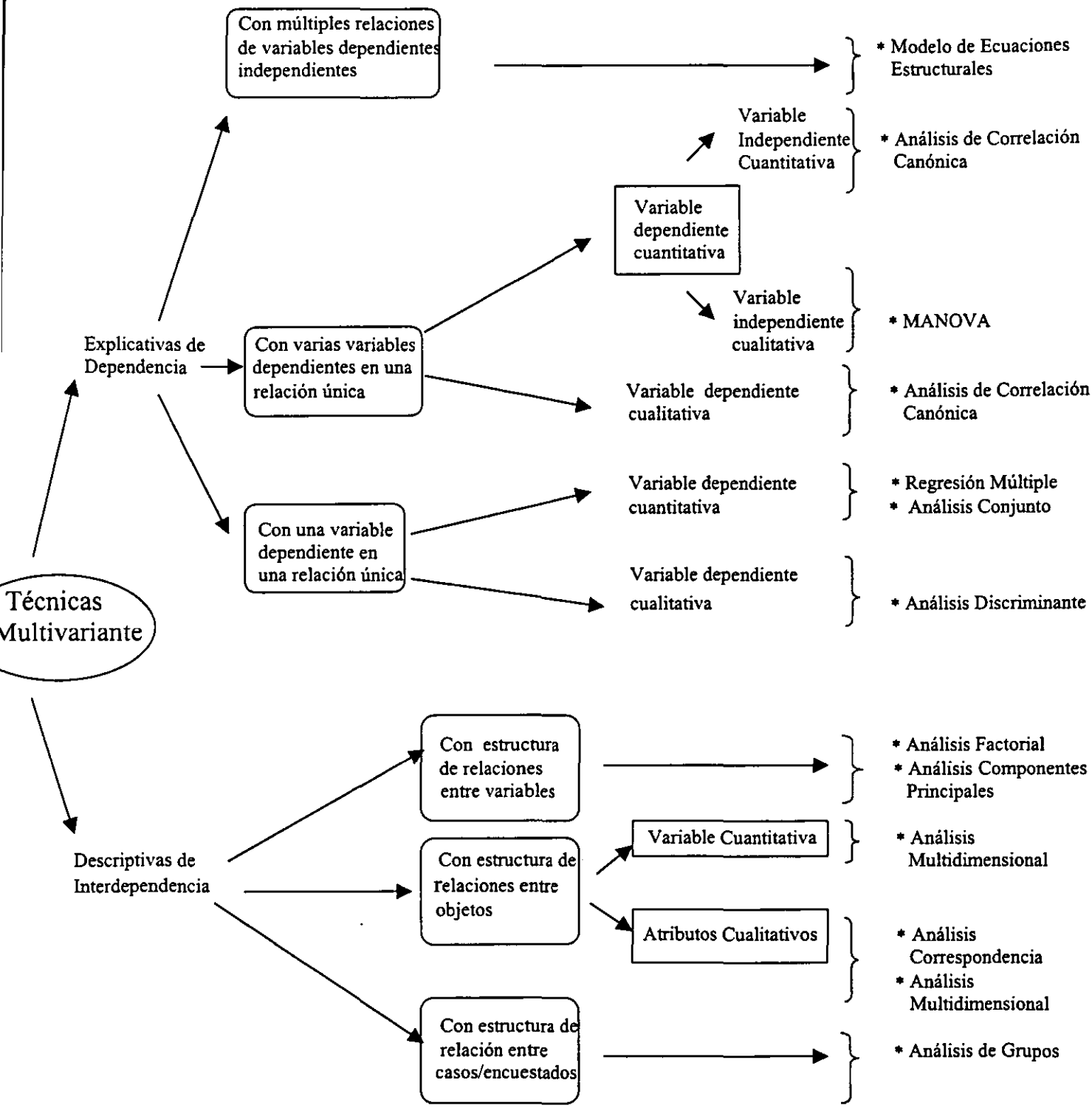


Figura 3.1. Clasificación de las Técnicas de Análisis Multivariante.

La normalidad multivariante es mucho más difícil de contrastar, aunque existen varios test para situaciones en que la técnica de análisis multivariante se ve particularmente afectada por la violación de este supuesto, generalmente se contrasta la normalidad univariante y aunque esta no garantiza la normalidad multivariante, si todas las variables cumplen ese requisito, entonces cualquier incumplimiento de este supuesto es generalmente insignificante. Las violaciones de este supuesto son además poco influyentes si los tamaños muestrales son grandes.

Supuesto de Homocedasticidad

Este supuesto está referido a las relaciones de dependencia entre variables.
Se supone que las variables dependientes tengan iguales niveles de varianza a lo largo del rango del predictor de las variables. La homocedasticidad es deseable porque la varianza de la variable dependiente no debe concentrarse solo en un limitado rango de los valores independientes.
El test de Levense se utiliza para evaluar si las varianzas de una única variable métrica son iguales a lo largo de cualquier cantidad de grupos. Para el contraste de más de una variable métrica se aplica el test de M de Box, válido para el análisis multivariante.

Supuesto de Linealidad.

Este es un supuesto implícito de todas las técnicas multivariante basadas en medidas de correlación incluyendo la regresión múltiple, el análisis logit, el análisis factorial y los modelos de ecuaciones estructurales. Es siempre aconsejable examinar todas las relaciones para identificar cualquier desplazamiento de la linealidad que pueda impactar la correlación..
La forma más común de evaluar la linealidad es examinar los gráficos de dispersión de las variables e identificar cualquier pauta no lineal en los datos.

CAPÍTULO 4.

Técnicas de Análisis Multivariante. Estimación de los Modelos.

En este capítulo se efectuará una descripción más detallada de los diferentes métodos tratados en este trabajo con el propósito de que se tenga una idea más amplia de las bases teóricas y conceptuales de los mismo, sin llegar a abordar en toda su amplitud los aspectos matemático–estadísticos que subyacen en cada técnica, ya que esto rebasaría las intenciones de este trabajo.

4.1. Técnicas descriptivas de interdependencia.

4.1.1. Técnicas Factoriales.

Técnicas Factorial es un nombre genérico que se le da a un conjunto de técnicas estadísticas multivariantes cuyo propósito principal es definir la estructura subyacente en una matriz de datos [Hair, 1999]. El análisis factorial proporciona la base para crear una nueva serie de variables que incorporan el carácter y naturaleza de las variables originales en una cantidad de nuevas variables más reducidas. De esta manera, se pueden reducir los problemas que se asocian con las grandes cantidades de variables. Dentro de estas técnicas se incluye el análisis de componentes principales, el análisis factorial común y el análisis de correspondencias.

Supuestos de las Técnicas Factoriales.
Desde un punto de vista estadístico se pueden obviar los supuestos de normalidad, homogeneidad y linealidad, siendo concientes de que su incumplimiento produce una disminución en las correlaciones observadas.
Debe asegurarse de que exista suficiente correlación en la matriz de datos para justificar la aplicación del análisis factorial.
Una manera de determinar la conveniencia del análisis factorial es examinar la matriz de correlación entera. El contraste de esfericidad de Bartlett, una prueba estadística para la presencia de correlación entre las variables es una de estas medidas.

Test de esfericidad de Bartlett: Este test plantea como hipótesis nula la existencia de incorrelación lineal entre las variables (poblacionales) lo que se traduce como la probabilidad estadística de que la matriz de correlación de las variables sea una matriz identidad .
Cuando la hipótesis nula se rechaza, los datos ponen en evidencia la existencia de correlaciones entre las variables.
Otra medida para cuantificar el grado de incorrelación entre las variables y la conveniencia de un análisis factorial es la medida de suficiencia de muestreo (MSA). Este índice se

extiende de 0 a 1, llegando a 1 cuando cada variable es perfectamente predicha sin error por las demás variables. Esta medida puede ser interpretada con las siguientes directrices:

0.80 o superior	Sobresaliente
0.70 – 0.79	Regular
0.60 – 0.69	Mediocre
0.50 – 0.59	Despreciable
por debajo de 0.50	Inaceptable

Se debe examinar primero los valores de MSA para cada variable y excluir aquellos que caen en la gama de inaceptables. Una vez que las variables individuales logran un valor aceptable, se puede valorar el MSA general y tomar una decisión sobre la continuidad del análisis factorial.

Validación en el Análisis factorial.
El método más directo de validación de los resultados consiste en adoptar una perspectiva de confirmación, valorando la replicabilidad de los resultados, bien dividiéndola muestra de datos originales, o bien con una muestra adicional.

4.1.1.1. Análisis de Componentes Principales.

Originalmente introducida por Pearson (1901) e independientemente por Hotelling (1933), la idea básica de este método es describir la variación de un conjunto de datos multivariantes en términos de un conjunto de variables incorrelacionadas (llamadas componentes o factores). Cada una de estas componentes es una combinación lineal de las p variables originales y el método empleado para su construcción garantiza que están ordenadas de acuerdo con la información que contienen, cuantificada a través de su varianza. [Everitt, 1991]
Esta técnica puede entenderse por tanto como un método de reducción de la dimensión, puesto que seleccionando las m ($m<p$) primeras componentes garantizamos que contienen un elevado porcentaje de la información de las variables originales [Pérez, 1997].

Estimación de los Factores.
Cálculo del número de factores: Para determinar cuantos factores se deben extraer, generalmente se comienza con algún criterio predeterminado, como puede ser el porcentaje de la varianza total acumulada extraído o el criterio de la raíz latente (cualquier factor individual debe justificar la varianza de por lo menos una variable.
Después de estimar la solución inicial se calculan varias soluciones de pruebas adicionales normalmente un factor menos que el número inicial y dos o tres factores más.
Examinando las matrices de factores respectivas se escoge el número de factores que representa mejor los datos. La matriz de factores contiene las cargas factoriales, que no son más que las correlaciones entre cada variable y el factor. Mientras más alta sea el valor absoluto de la carga factorial más importante es.

El primer factor puede interpretarse como el mejor resumen de las relaciones lineales manifestadas por los datos. El segundo factor se define como la segunda mejor combinación lineal de las variables sujetas a la restricción de que sea ortogonal al primer factor, o sea , que se deriva de la varianza restante después de la extracción del primer factor. Los factores subsiguientes se definen de forma análoga, hasta haber agotado la varianza de los datos.
Una herramienta importante al interpretar los factores es la rotación de factores, para lo cual se giran en el origen los ejes de referencia de los factores hasta alcanzar una determinada posición.
La rotación es deseable porque simplifica la estructura de los factores y proporciona una solución teórica más simplificada, además en muchos casos la rotación de los factores mejora la interpretación disminuyendo posibles ambigüedades de la solución no rotada.

Métodos de rotación: En la práctica todos los métodos de rotación tienen el objetivo de simplificar las filas o columnas de la matriz de factores para facilitar la interpretación. En una matriz de factores, las columnas representan los factores y las filas se corresponden con las cargas de las variables para cada factor.
Los métodos de rotación más frecuentemente utilizados son los de rotación ortogonal, entre ellos el QUARTIMAX (simplifica las filas de la matriz) y el VARIMAX (se centra en la simplificación de las columnas de la matriz de factores). En la figura 3.1 se puede ver un ejemplo de rotación ortogonal [Hair, 1999] de la inspección visual se puede apreciar que dos grupos de variables 1 y 2 van juntas y 3, 4 y 5 conforman otro grupo. Sin embargo este patrón de variables no es tan obvio a partir de las cargas de los factores no rotados.

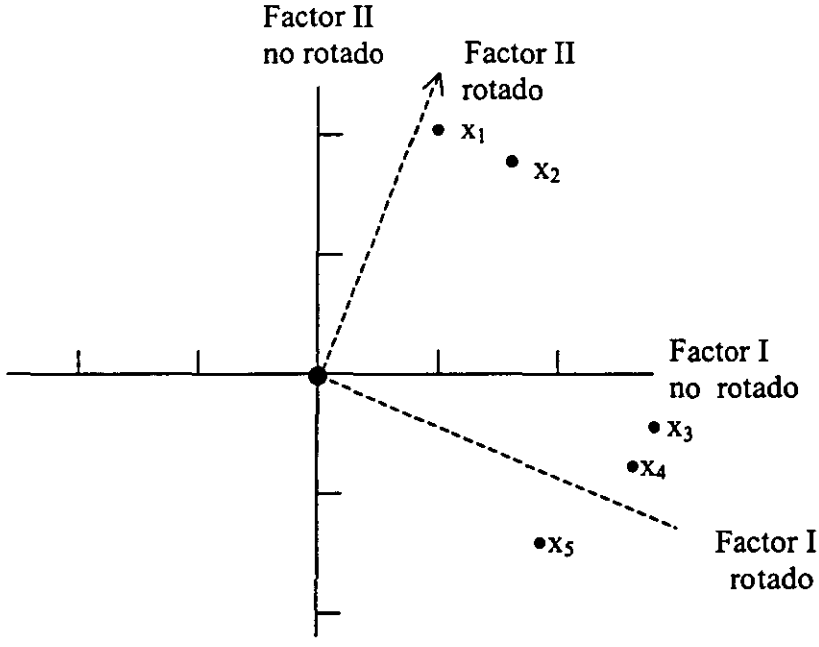


Figura 4.1 Ejemplo de rotación factorial ortogonal.

Al interpretar los factores es necesario decidir que cargas factoriales deben ser consideradas. Desde un punto de vista práctico las cargas factoriales mayores a ± 0.30 se

consideran en el nivel mínimo, las cargas de ± 0.40 se consideran más importante y las cargas de ± 0.50 o mayores, se consideran prácticamente significativas(criterio válido para muestras que superan las 100 observaciones)
Otro criterio basado en la significación estadística, utiliza el concepto de potencia estadística para especificar cargas factoriales consideradas significativas según diferentes tamaños muestrales (ver tabla 4.1)

Carga factorial	Tamaño muestral necesario para la significación
0.30	350
0.35	250
0.40	200
0.45	150
0.50	120
0.55	100
0.60	85
0.65	70
0.70	60
0.75*	50

Tabla 4.1. Criterio para la selección de cargas factoriales significativas a partir del tamaño muestral.

*cargas superiores al 75% explicarían más del 50% de la varianza y se consideran cargas extremadamente elevada fuera de lo normal.

Comunalidad: Después de agrupada cada variable en sus respectivos factores debe valorarse la comunalidad. La comunalidad representa la proporción de varianza con la que contribuye cada variable a la solución final. En este sentido debe evaluarse si alcanza niveles aceptables de explicación.

4.1.1.2. Análisis Factorial Común.

El análisis factorial común (AFC) parte de una matriz de correlaciones entre p variables y de la hipótesis de que dichas correlaciones no son aleatorias, sino que se deben a que las variables comparten causas comunes, llamadas factores.

El objetivo del AFC es precisamente identificar dichos factores comunes y cuantificarlos, para lo que se formula un modelo que requiere hipótesis estadísticas y la aplicación de métodos inferenciales.
La estimación de los factores y las contribuciones de cada variable a los factores (denominadas cargas de los factores) constituye todo lo que hace falta para el análisis.
El proceso es similar al descrito para el ACP, la diferencia entre estos dos métodos radica en la especificación de la matriz factorial (método para extraer los factores). El ACP utiliza la varianza total y el AFC se basa en la varianza común, desestimando la varianza de cada variable específica y la varianza de error (por poca fiabilidad en el proceso de recolección de datos, por errores de medición o componentes aleatorios en el fenómeno medido).
Se utiliza el ACP cuando el objetivo es resumir la mayoría de la información original en una cantidad mínima de factores. Por el contrario se utiliza el AFC para identificar los factores subyacentes o las dimensiones que reflejan que es lo que las variables comparten en común.
Para cualquiera de los dos métodos es necesario determinar el número de factores que representan la serie de variables originales.

4.1.1.3. Análisis de Correspondencias.

Las técnicas factoriales anteriormente expuestas se basan en una matriz *nxp* a partir de la cual se calculan las matrices de covarianza, correlaciones, comunalidades, etc. Estos métodos son válidos solo para variables cuantitativas, cuando partimos de variables cualitativas no disponemos de una matriz de datos, sino, en todo caso de una tabla de contingencia (frecuencias) , multidimensional si nos referimos a tres o más variables.
El análisis de correspondencia ofrece, por lo tanto, una representación multivariante de interdependencia para datos no métricos que no es posible realizar con otros métodos.
Si nos limitamos a dos variables cualitativas la técnica se denomina *Análisis de Correspondencia Simple*, mientras que si el número de ellas es superior, se dice que se trata de un *Análisis de Correspondencia Múltiple.*

Suponiendo que tenemos dos variables cualitativas x , y clasificadas en dos modalidades r y s modalidades respectivamente, esta información puede ser resumida en una tabla de frecuencias absolutas o relativas a partir de la cual se inicia el análisis de correspondencia.

	y_1	...	y_s	f_i
x_1	f_{11}	...	f_{1s}	f_1
...	...	...	...	...
x_r	f_{1r}	...	f_{rs}	f_r
f_j	f_1	...	f_s	1

Tabla 4.2 Tabla de frecuencias

La comparación de las modalidades de x o y no pueden hacerse directamente sobre esta tabla, pues la distribución de frecuencia relativa de dos categorías puede ser muy similar y en cambio existir disparidad en la frecuencia total entre ellas.
Por este motivo debemos definir funciones condicionadas que permitan comparar la distribución de categorías con independencia de las frecuencias marginales.

El modelo de correspondencia es simétrico en el sentido de que filas y columnas juegan el mismo papel.
Dado que los perfiles filas y columnas pueden tener una dimensión considerable (r y s elevados), nuestro objetivo será reducir esta dimensión, es decir poder representar las filas y las columnas en un espacio de dimensión menor (generalmente dos) de modo que se reproduzcan las distancias entre los perfiles originales de la forma más fiel posible.
Para estudiar la similitud entre dos filas, se define la distancia χ^2 entre ellas como la distancia euclideana entre sus perfiles, ponderando inversamente por las frecuencias marginales de las columnas (de forma análoga se define la distancia entre las columnas).
Al igual que en el ACP el interés de realizar el proceso venía justificado por la existencia de correlación lineal entre las variables, en este caso al aplicar un análisis de correspondencia se parte del hecho de haber detectado cierto grado de asociación entre los atributos.
Si abordamos el estudio basándonos en los perfiles fila, la matriz que se diagoniza contiene las asociaciones entre las categorías columnas, y su traza se identifica con la inercia total de los datos de partida.
El criterio para determinar el primer eje es que maximice la inercia proyectada sobre él y la restricción será que su norma sea unitaria.
El resultado de la extracción de factores es el mismo independientemente de que hagamos el desarrollo basándonos en las filas o en las columnas. En definitiva, se tratará de aplicar un análisis de componentes principales a cualquiera de las dos matrices. De esta forma se consigue cuantificar la información cualitativa inicial. Ahora disponemos de unas variables cuantitativas que pueden utilizarse en análisis posteriores.
Será preciso además dar una interpretación a los factores y estudiar la bondad de la representación de cada categoría. Para ello disponemos de dos instrumentos: las contribuciones absolutas y las contribuciones relativas.
Las contribuciones absolutas son el porcentaje de inercia de un factor imputable a cada categoría de una fila (de igual forma se define para una columna). Mediante las contribuciones absolutas a un factor podemos saber que categorías tienen más peso en un factor.
Las contribuciones relativas de los factores a una categoría tienen una interpretación análoga a las comunalidades en el ACP. Indican la proporción de inercia de cada categoría explicada por el factor y puede entenderse como correlaciones al cuadrado entre el factor y la categoría. La suma de las contribuciones relativas de los factores retenidos a cada categoría podemos interpretarla como una medida de la calidad de la representación de la categoría en los nuevos ejes.
Los factores relacionan simultáneamente filas y columnas en un único gráfico conjunto. El resultado es una representación de categorías de filas y/o columnas en el mismo gráfico.
Una vez establecida la dimensionalidad se puede identificar una asociación de categorías con otras categorías por su proximidad.

4.1.2. Análisis de Grupos (*Cluster*).

El análisis de grupos ó análisis cluster es un método multivariante de clasificación de datos, cuyo objetivo es formar grupos de individuos homogéneos respecto a un conjunto de características, que pueden ser cualitativas o cuantitativas.

La esencia del análisis cluster es encontrar grupos de cosas (objetos, unidades experimentales, variables, etc.) tal que las cosas dentro de un grupo sean más similares entre sí (en algún sentido indicado por su medición) que las cosas incluidas en los otros grupos. [Gnanadesikan, 1997]
Existen diversos algoritmos cluster y dependiendo del tipo de datos manejados y de los objetivos concretos del estudio se aplicarán unos u otros.

Supuestos del Análisis de Grupos.
El análisis de grupos es una técnica de inferencia estadística en la que se analizan los parámetros de una muestra en la medida en que puedan ser representativos de una población por lo que no tiene fundamentos estadísticos. Las exigencias de normalidad, homocedasticidad y linealidad que son importantes en otras técnicas tienen poco peso en el análisis cluster.

La información de partida de este análisis es una matriz (*nxp*) que contiene las observaciones de las *p* variables clasificadoras para los *n* individuos que se desean agrupar

En el proceso de aplicación del análisis cluster debemos preocuparnos por la forma en la que mediremos la separación entre los valores de los distintos individuos, con arreglo a qué criterios se formarán los grupos de nivel básico y cómo iremos reagrupando éstos en otros de orden superior; por tanto, será preciso tomar decisiones sobre los tres aspectos siguientes:

a) Medida de distancia
b) Método de formación de grupos
c) Criterio para combinar grupos

Medidas de Distancias.
Cada grupo o cluster estará constituido por un conjunto de individuos similares por lo que será necesario calcular en una primera etapa las similaridades o de forma equivalente y en sentido opuesto, las distancias entre los *n* individuos. La medida de distancia más utilizada cuando los datos son cuantitativos es la euclideana al cuadrado, en cuyo caso, si representamos por x_i el vector de coordenadas del i-ésimo individuo (valores asociados al individuo para todas las variables), la distancia entre los individuos *i* e *i'* podemos expresarla en forma vectorial como:

$$d^2(i,i') = (x_i - x_{i'})'(x_i - x_{i'})$$

En el caso de datos cualitativos puede emplearse la distancia χ^2
En función del tipo de datos disponibles existe una gran variedad de medidas posibles, cada una con sus ventajas e inconvenientes (podemos citar a modo de ejemplo las basadas en coeficientes de correlación como el de Pearson o el de Kendall para datos ordinales o en el número de concordancias para el caso de variables dicotómicas de ausencia-presencia, etc.)

Un problema que se presentará con frecuencia es que las variables vengan expresadas en unidades diferentes. En el análisis cluster todas las variables tienen la misma importancia,

por lo que será conveniente trabajar sobre las variables tipificadas, debiendo tenerse en cuenta que este proceso tiende a diluir las diferencias entre los grupos.

También puede presentarse el problema de que el número de variables clasificadoras sea muy grande o que éstas estén correlacionadas. Ambos problemas pueden resolverse aplicando previamente un análisis factorial con lo que habremos reducido el número de variables y además garantizaremos que éstas serán incorrelacionadas. Otra alternativa para superar al mismo tiempo el problema de la correlación y de las unidades es utilizar la distancia de Mahalanobis.

Métodos de Formación de Grupos.

En cuanto a los métodos de formación de grupos podemos distinguir entre el cluster jerárquico y el no jerárquico. Los métodos jerárquicos configuran grupos con estructura de árbol, de forma que los cluster de niveles más bajos van siendo englobados en otros de niveles superiores. Los métodos no jerárquicos o de partición asignan los individuos a grupos diferenciados que el propio análisis configura, sin que unos dependan de otros; es decir, no habrá una estructura vertical de dependencia entre los grupos formados. En este caso es preciso indicar a priori el número de cluster que se desea formar.

Los métodos jerárquicos proporcionan información más amplia que los no jerárquicos, pero presentan el inconveniente de que requieren un elevado número de cálculos, Por ello, cuando se dispone de un gran número de datos suelen emplearse al menos en una primera etapa los métodos no jerárquicos y posteriormente con los cluster así formados puede pasarse a aplicar un algoritmo jerárquico.

Cluster jerárquico

Dentro de los métodos jerárquicos pueden emplearse dos tipos de algoritmos para la formación de grupos: aglomerativo o ascendente y divisivo o descendente. El primero de ellos comienza con tantos cluster como individuos y en cada etapa se forma un grupo por unión de dos individuos aislados, de dos grupos o de un individuo con un grupo formado en una etapa anterior; el final del proceso es un grupo único formado por todos los individuos.

Entre los criterios alternativos empleados en los métodos jerárquicos para combinar grupos podemos citar los siguientes:

Enlace simple: toma como distancia entre dos grupos la existente entre los componentes de ambos más próximos.

Enlace completo: toma como distancia entre los grupos la correspondiente a los elementos más distantes de ambos.

Enlace centroide: la distancia entre dos grupos será la que media entre sus centros de gravedad respectivos.

Enlace promedio: considera como distancia entre dos grupos la distancia media entre todos los pares posibles de casos (uno de cada cluster).

Enlace de mínima varianza (Ward): se trata de efectuar la agregación de modo que el aumento de la variabilidad dentro de los grupos sea lo más pequeño posible.

Los resultados de la aplicación de un algoritmo cluster pueden variar sensiblemente según la elección del método de enlace y de la medida de distancia. Por otra parte, los métodos jerárquicos presentan la desventaja frente a los no jerárquicos de que los individuos mal clasificado en una fase inicial no pueden ser reclasificados.
Como resultado de aplicar el algoritmo se obtiene una representación gráfica en forma de árbol invertido denominada dendograma que muestra las etapas de formación de los grupos En el dendograma es posible conocer la composición de los grupos según el número de ellos que hayamos establecido solo trazando una vertical a la altura correspondiente, y además la longitud de las barras indica la distancia entre los grupos que se combinan. (ver ejemplo figura 4.2)

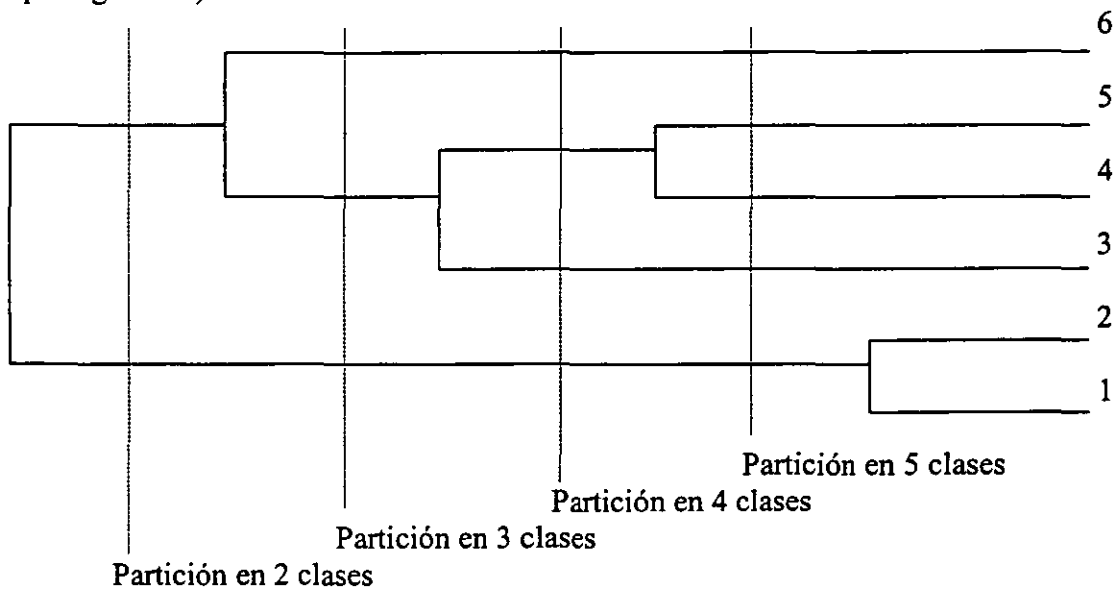


Figura 4.2. Ejemplo sencillo de dendograma para 6 observaciones.

Cluster no jerárquico.
Entre los algoritmos más usados de este tipo está el de agregación alrededor de centros móviles. Se trata de un proceso iterativo en el que los individuos se agrupan en cada etapa, cambiando los centros de gravedad de los grupos.
Este método se basa en una idea muy simple; conseguir que la dispersión de los grupos sea lo más pequeña posible.
En el caso de una única variable clasificadora ($p = 1$) y teniendo en cuenta la descomposición usual de análisis de la varianza: VT = VE + VNE, el criterio anterior supone minimizar la variación dentro de los grupos VNE o de forma equivalente maximizar la variación entre grupos VE, dado que la variabilidad total es constante.
En el caso general de p variables se partirá de la descomposición de la matriz T de orden p, que recoge la variabilidad total como T = F + W donde la matriz F representa la variabilidad entre grupos o factorial y W la variabilidad dentro de los grupos o residual. El criterio de agrupación consistirá entonces en maximizar la traza de la matriz W.
Este procedimiento tiene una interpretación intuitiva muy clara ya que equivale a minimizar la suma de las distancias al cuadrado entre cada individuo y el centro de gravedad del cluster al que es asignado.

Fijados los centros de los *k* cluster que deseamos formar, asignaremos cada individuo al grupo cuyo centro se encuentra más próximo.
Tras finalizar la primera asignación se comprueba si se verifica el criterio de convergencia fijado, un criterio posible consiste en detener el proceso cuando la distancia entre los nuevos centros de gravedad no haya aumentado sensiblemente en esa etapa. En caso negativo, se realiza una nueva iteración teniendo en cuenta los nuevos centros de los cluster, y así se continúa hasta que se satisfaga el criterio de parada.
Como en este método es necesario fijar el número de cluster a priori habitualmente se repite la aplicación del algoritmo para distintos valores de *k* para así poder elegir la clasificación que mejor se ajuste al objetivo del problema, o bien la de interpretación más clara.
Una vez finalizado el proceso podemos utilizar algunos de los contrastes para comprobar la adecuación de la clasificación . Así bajo los supuestos de normalidad e igualdad de varianza entre los cluster, podemos considerar un ANOVA utilizando los contrastes F entre variabilidad explicada y no explicada (en este caso la hipótesis nula será que las medias en los cluster finales son iguales).

La interpretación de los resultados del análisis cluster implica el examen de cada grupo , asignando una etiqueta precisa que describa su naturaleza.

Validación.
Aunque no existe un método único para asegurar la validez y relevancia práctica de la solución cluster, se han propuesto diferentes aproximaciones. Entre ellas tenemos:
1.- Dividir la muestra en dos grupos, analizarlas por separado y comparar los resultados.
2.- Una forma modificada de división de la muestra empleando los centros de los grupos obtenidos desde una solución cluster para definir conglomerados a partir de otras observaciones para comparar después los resultados
3.- Una forma directa de validación cruzada.

4.1.3. Análisis Multidimensional.

El análisis multidimensional (también conocido como elaboración de mapas preceptúales) consiste en un conjunto de procedimientos que ayudan a identificar las dimensiones subyacentes claves en las evaluaciones de los objetos de estudio, utilizadas por los encuestados. Es una técnica que permite al investigador determinar la imagen percibida relativa de un conjunto de objetos (empresas, productos, ideas u otros) [Hair, 1999]
Una vez que se dispone de los datos el análisis multidimensional puede ayudar a determinar que dimensiones utilizan los encuestados cuando evalúan los objetos, cuantas dimensiones pueden utilizar en una situación particular, la importancia relativa de cada dimensión y como se relacionan los objetos.
El propósito del análisis multidimensional es transformar los juicios de similitud o preferencia del consumidor en distancias representadas en un espacio multidimensional. El mapa perceptual resultante muestra la situación relativa de todos los objetos.
En el análisis multidimensional cada encuestado proporciona evaluaciones de todos los objetos que se están considerando, por lo que se puede obtener una solución para cada individuo. El énfasis no se pone en los objetos, sino en como el individuo percibe los objetos.

La elaboración de mapas preceptúales puede desarrollarse con técnicas de composición o descomposición en función de la naturaleza de las respuestas obtenidas del individuo en relación con el objeto.
El método de descomposición mide solo la impresión o evaluación conjunta de un objeto y a continuación intenta obtener posiciones espaciales en un espacio multidimensional que refleje estas percepciones.
El método de composición es una aproximación alternativa que emplea varias técnicas de análisis multivariante que se usan en la formación de una impresión o evaluación basada en una combinación de atributos específicos.

Supuestos.
Los supuestos del análisis multidimensional se centran principalmente en la comparabilidad y representatividad de los objetos que están siendo evaluados y de los encuestados.
Aunque el análisis multidimensional no tiene supuestos restrictivos en cuanto a tipos de datos o forma de la relación entre las variables, si requiere del cumplimiento de los siguientes principios.

- Variación en la dimensionalidad: Cada encuestado no percibirá la misma dimensionalidad en un estímulo.
- Variación en importancia: Los encuestados no necesitan asignar el mismo nivel de importancia a una dimensión, incluso si todos ellos perciben esta dimensión.
- Variación en el tiempo: Los juicios de un estímulo en términos tanto de dimensiones o niveles de importancia no tienen porque permanecer estables en el tiempo, o lo que es lo mismo, no puede esperarse que los encuestados mantengan las mismas percepciones durante largos períodos de tiempo.

Posicionamiento de un objeto en un mapa perceptual.
Los programas de análisis multidimensional siguen un proceso común de determinación de posiciones óptimas que puede resumirse en cuatro pasos:

1.- Selección de una configuración inicial de estímulos (S_k) respecto a una dimensión (t) deseada. Existen varias opciones para obtener las configuraciones iniciales, las dos más ampliamente utilizadas son configuraciones, o bien aplicadas por el investigador basándose en datos previos, o generadas por puntos pseudo aleatorios desde una distribución multivariante aproximadamente normal.

2.- Calcular las distancias entre los puntos de estímulo y comparar las relaciones con una medida de ajuste, normalmente una medida de estrés.
3.- Si la medida de ajuste no llega a un valor límite seleccionado, hay que encontrar una nueva configuración para la cual la medida de ajuste se minimice aún más.
4.- Una vez que se ha alcanzado una medida de ajuste satisfactoria, la dimensionalidad se reduce a uno y el proceso se repite hasta que se alcance la menor dimensionalidad con una medida aceptable del ajuste.

La medida de estrés es simplemente una medida de lo bien (o mal) que las distancias representadas en un mapa concuerdan con las clasificaciones dadas por los encuestados.

Puntos ideales.
Un punto ideal (PI) es aquel que representa la combinación más preferida de atributos percibidos. La posición de este punto ideal, en relación con otros puntos en el mapa perceptual derivado, define preferencias relativas de tal forma que los puntos que están lejos de este ideal serán menos preferidos. El punto ideal se sitúa de tal forma que la distancia de cualquier objeto a este punto expresa cambios en la preferencia. Por ejemplo a partir de lo reflejado en la figura 4.3 uno puede suponer que el orden de preferencia de la persona es *C, F, D, E, A, B*.

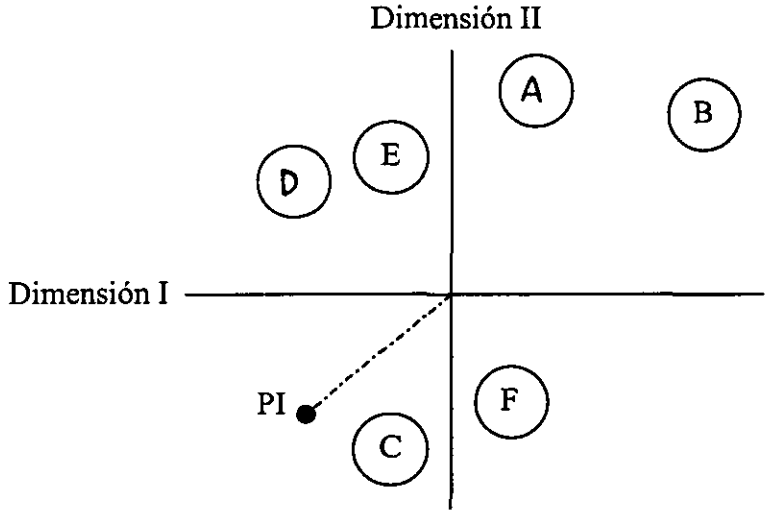


Figura 4.3 Punto ideal de un encuestado dentro de un mapa perceptual bidimensional.

Muchos encuestados con puntos ideales en la misma área general, representa un mercado potencial de segmento de personas con preferencias similares.

Interpretación de los resultados.
Una vez obtenido el mapa perceptual la interpretación depende del enfoque utilizado en su elaboración.
Enfoque de descomposición: El asunto más importante es la descripción de las dimensiones preceptúales y su correspondencia con los atributos (ya sea subjetiva u objetiva)
Enfoque de composición: El mapa de percepciones debe ser validado con otras medidas de percepción dado que las posiciones están totalmente definidas por los atributos especificados por el investigador.

Validación de los resultados.
La validación en el análisis multidimensional debido a su naturaleza altamente inferencial debe dirigirse a asegurar la generalidad de los resultados tanto en los objetos como en la población.
El método que se tiene con este propósito es la comparación de varias muestras, bien dividiendo la muestra original o recogiendo una nueva muestra. A menudo la comparación entre resultados se hace visualmente o con una correlación simple de resultados.

4.2. Técnicas Explicativas de Dependencias.

4.2.1. Análisis de Regresión Múltiple.

El análisis de regresión múltiple es una técnica estadística general utilizada para analizar relaciones entre una única variable dependiente métrica y varias variables independientes también métricas, su formula básica es:

$$\underset{\text{(métrica)}}{Y_1} = \underset{\text{(métricas)}}{X_1 + X_2 + X_3 + \ldots + X_n}$$

El objetivo de la regresión múltiple es usar las variables independientes, cuyos valores son conocidos, para predecir la única variable dependiente seleccionada por el investigador.
La creciente aplicación de la regresión múltiple se agrupa en dos amplias clases de problemas de investigación: predicción y explicación., problemas que no son mutuamente excluyentes.
Un ejemplo de una variante de análisis de regresión en su carácter predictivo es el análisis de series temporales, en el cual el único propósito es predecir y la interpretación de los resultados es útil solo como un medio de incrementar la precisión predictiva.
La regresión múltiple, en su variante explicativa, proporciona un medio de evaluar el grado y carácter de la relación entre las variables independientes y la dependiente.
El carácter de la regresión múltiple que la diferencia de su contrapartida univariante es la evaluación simultanea de las relaciones entre cada variable independiente y las medidas de la dependiente. Al realizar esta evaluación simultanea se determina la importancia relativa de cada predictor.
Finalmente la regresión múltiple proporciona también una idea de las relaciones entre las variables independientes en sus predicciones de la variable dependiente.
Estas interpretaciones son importantes dado que la correlación entre las variables independientes puede hacer que algunas variables sean redundantes es su esfuerzo predictivo, o sea que no sean necesarias para producir una predicción óptima. No se trata de reflejar sus relaciones individuales con la variable dependiente sino que indica que en un contexto multivariante, no son necesarias si se emplea otro conjunto de variables independientes.
Colinealidad es la asociación, medida como correlación, entre dos variables independientes.
Multicolinealidad se refiere a la correlación entre tres o más variables.
El impacto de la multicolinealidad consiste en reducir el poder predictivo de cualquier variable independiente individual en la medida en que está asociada con las otras variables independientes. Conforme aumenta la colinealidad,a varianza única explicada por cada variable independiente se reduce y el porcentaje de predicción compartida aumenta.
Para maximizar la predicción de un número específico de variables independientes, deben buscarse otras variables independientes que tengan una multicolinealidad baja con las otras variables independientes pero que también tengan correlaciones altas con la variable dependiente.

Supuestos básicos en el Análisis de Regresión Múltiple.
El concepto de correlación está basado en una relación lineal, siendo por lo tanto la linealidad del fenómeno medido un supuesto crítico del análisis de regresión.

La presencia de varianzas desiguales (heterocedasticidad) es uno de los supuestos que se incumplen habitualmente. El diagnóstico se puede hacer mediante gráficos de residuos o test estadísticos simples.
Otro supuesto que debe verificarse es el de independencia del término de error. Para identificar este hecho se utiliza el gráfico de residuos respecto a cualquier posible varianza secuencial. Si los residuos son independientes la forma puede parecer aleatoria y similar al gráfico de no correlación de los residuos.
El supuesto que con mayor frecuencia es violado es el de la normalidad de las variables independientes o dependiente, o ambas. El diagnóstico más simple es un histograma de los residuos donde se puede comprobar visualmente si la distribución se aproxima a la normal. Otro método más aconsejable es utilizar los gráficos de probabilidad normal.

Estimación del modelo de regresión.
En la mayoría de los casos de regresión múltiple se tiene un número posible de variables independientes entre las cuales elegir para incluirlas en la ecuación de regresión. Si el conjunto de variables está totalmente definido el análisis se utiliza en una aproximación confirmatoria. Si se desea elegir entre el conjunto de variables independientes se pueden emplear varias aproximaciones para buscar el mejor modelo de regresión (métodos de búsqueda secuencial como la estimación por etapas, la eliminación progresiva y regresiva y procesos combinatorios).
Los métodos de búsqueda secuencial tienen en común la aproximación general de estimación de las ecuaciones de regresión con un conjunto de variables y a continuación añadir o eliminar selectivamente variables hasta que se consiga alguna medida criterio conjunto.
Los métodos combinatorios son fundamentalmente un proceso de búsqueda generalizada a lo largo de todas las combinaciones posibles de variables independientes.

Ecuación de Regresión.
Para predecir la variable dependiente se aplica la ecuación de regresión múltiple:

$$y = b_o + b_1x_1 + b_2x_2 + \ldots + b_kx_k \ ,$$

donde b_o es una constante de regresión,; b_1 , b_2 , ... , b_k son los valores que indican el cambio en la variable dependiente ante una cambio unitario en la variable independiente respectiva (coeficientes de regresión) y x_1 , x_2 , ... , x_k , son los valores de las variables independientes a partir de las cuales se hace la predicción de la variable dependiente y.
Una información básica que proporciona la regresión múltiple es el coeficiente de correlación múltiple (R), que representa la correlación entre la variable dependiente y todas las variables independientes tomadas en conjunto, el coeficiente varía entre 0 y 1.00 y mientras mayor sea su valor, las variables independientes explican en mayor medida la variación en la variable dependiente. Sin embargo la fuerza de la relación se representa mejor por R^2 que nos indica el porcentaje de variación en la variable dependiente debido a las variables independientes.
Como parte de todos los programas de regresión se da un coeficiente de regresión ajustado (R^2 *ajustado*) que se interpreta igual que R^2 . El R^2 *ajustado* se hace más pequeño a medida que tenemos menos observaciones por variables independientes y es particularmente útil

para comparar las diferentes ecuaciones de regresión estimadas con distintas variables independientes o diferentes tamaños muestrales, dado que marca límites para el número específico de variables independientes y para el tamaño muestral sobre el que se basa cada modelo.
Cuando los valores de las variables independientes han sido estandarizados antes de estimar la ecuación de regresión los coeficientes de regresión obtenidos se denominan coeficientes *beta* que indican el impacto relativo sobre la variable dependiente de un cambio en una desviación estándar de cada variable o dicho en otra forma, indica el peso o influencia que tiene cada variable independiente sobre la dependiente.

Validación de los resultados.
La aproximación más apropiada para la validación es contrastar el modelo de regresión mediante la extracción de una nueva muestra de la población. Cuando esto no sea posible se puede dividir la muestra original en dos partes y utilizar una submuestra para crear el modelo y la otra para validación para contrastar la ecuación.
Una alternativa a la obtención de muestras adicionales es también emplear la muestra original de forma especializada mediante el cálculo del estadístico PRESS, una medida similar a R^2 utilizada para evaluar la precisión predictiva del modelo de regresión estimado.

4.2.2. Análisis Discriminante.

El análisis discriminante es una técnica multivariante basada en el estudio de las características diferenciales de una serie de grupos definidos a priori [Pérez, 1997]. El propósito básico del análisis discriminante es estimar la relación entre *una única variable dependiente* no métrica y un conjunto de *variables independientes métricas*, de la forma general:

$$\underset{\text{(no métrica)}}{Y_1} = \underset{\text{(métricas)}}{X_1 + X_2 + X_3 + \ldots + X_n}$$

Mediante esta técnica es posible alcanzar los siguientes objetivos:

- Explicar la pertenencia de los individuos de una muestra a uno de los grupos, en función de un conjunto de p variables, cuantificando la importancia relativa de cada una de ellas.
- Predecir a que grupo pertenece un individuo que no forma parte de los datos analizados para el que se conoce el valor de las p variables, pero no el grupo al que pertenece.

En muchos casos la variable dependiente consta de dos grupos o clasificaciones, por ejemplo masculino frente a femenino; en otras ocasiones se incluyen más de dos datos, como por ejemplo uno de tres grupos que comprenda la clasificación en alto, mediano y bajo. Cuando se incluyen dos clasificaciones la técnica es conocida como análisis discriminante de dos grupos . Cuando se identifican tres o más grupos, la técnica es conocida como análisis discriminante múltiple.
Como vemos el número de grupos de la variable dependiente puede se de dos o más pero estos grupos deben ser mutuamente excluyentes y exhaustivos.
El análisis discriminante implica obtener una combinación lineal de dos o más variables independientes que discrimine de la mejor forma los grupos definidos a priori. La

combinación lineal para el análisis discriminante es también conocida como función discriminante.

Supuestos de análisis discriminante.
Los supuestos claves para obtener la función discriminante son el de normalidad multivariante de las variables independientes y el de estructura (matrices de covarianza-dispersión) desconocidas pero iguales para los grupos. Si los supuestos no se cumplen debe buscarse un métodos alternativo.
Otra característica de los datos que puede afectar el proceso de clasificación es la multicolinealidad de las variables independientes.
Un supuesto implícito es que todas las relaciones son lineales. Las relaciones no lineales no están representadas en la función discriminante a menos que se realicen transformaciones específicas de las variables para representar los efectos no lineales.
Finalmente los casos atípicos pueden tener una influencia sustancial en la precisión clasificatoria de cualquier resultado por lo que se deben eliminar si es necesario.

Estimación del modelo de Análisis Discriminante.
Para la obtención de la función discriminante debe definirse primeramente el método de estimación a emplear. Se pueden utilizar dos métodos de cálculo para derivar una función discriminadora: el *método simultaneo* y el *método por etapas.*
La estimación simultanea implica un cálculo de la función discriminante donde todas las variables son consideradas simultáneamente. Este método es apropiado cuando se quieren introducir todas las variables independientes en el análisis y no se está interesado en observar resultados intermedios basados solamente en variables que discriminan mejor.
La estimación por etapas es una alternativa al enfoque simultaneo. Incluye las variables independientes dentro de la función discriminante de una en una, según su capacidad discriminatoria.
El enfoque por etapas comienza eligiendo la variable que mejor discrimina. La variable inicial se empareja con cada una de las variables independientes, de una en una, y se elige la que más consigue incrementar la capacidad discriminante de la función en combinación con la primera variable. Las restantes variables se seleccionan de forma similar.
Algunas variables seleccionadas previamente pueden ser eliminadas si la información que contiene sobre las diferencias del grupo está contenida en alguna combinación de otras variables incluidas en posteriores etapas. Al final o bien todas las variables han sido incluidas en la función o se habrá considerado que las variables excluidas no contribuyen significativamente a una mejor discriminación.
Una vez identificadas las funciones discriminantes significativas la atención se desplaza a averiguar el ajuste global de las funciones consideradas, lo que implica: calcular la puntuación Z para cada observación, evaluar diferencias de grupos sobre la puntuación Z discriminante y valorar la precisión en la predicción de pertenencia al grupo.
La puntuación Z es una medida métrica que permite comparar observaciones para cada función.
Una medida resumen de la diferencia entre grupos es una comparación de los centroides de grupo, que no es más que el promedio de las puntuaciones discriminantes (Z) para todos los individuos dentro de un grupo particular.

Los centroides indican la situación más común de cualquier individuo de un determinado grupo, y una comparación de los centroides de los grupos muestra lo apartado que se encuentran los grupos a lo largo de la dimensión que se está contrastando.
La capacidad predictiva de la función discriminante se mide con el rango de aciertos, el cual se obtiene de la matriz de clasificación de cada individuo (matriz que contiene la clasificación de cada individuo dentro de cada grupo a partir de su puntuación Z discriminante)
El criterio para considerar como aceptable el nivel de capacidad predictiva para una función discriminante puede ser que supere el porcentaje que podría ser clasificado correctamente de forma aleatoria (sin la ayuda de la función discriminadora).

Interpretación de los resultados.
Dentro de este proceso se examinan las funciones discriminantes para determinar la importancia relativa de cada variable independiente en la discriminación de los grupos.

Validación de los resultados.
Este paso se puede realizar mediante división de la muestra original o validación cruzada pero es posible emplear una muestra adicional como la ampliación de la muestra.

4.2.3. Análisis Multivariante de la Varianza.

El análisis multivariante de la varianza (MANOVA) es una técnica de dependencia que mide las diferencias de *dos o más variables dependientes métricas* basadas en un *conjunto de variables independientes no métricas* [Hair, 1999]. El MANOVA puede ser representado según la siguiente expresión general:

$$\underset{\text{(métricas)}}{Y_1 + Y_2 + Y_3 + \ldots + Y_n} = \underset{\text{(no métricas)}}{X_1 + X_2 + X_3 + \ldots + X_n}$$

El MANOVA se clasifica como proceso multivariante debido a que se utiliza para valorar diferencias entre grupos a través de múltiples variables dependientes métricas de forma simultanea. Es particularmente útil cuando se utiliza conjuntamente con diseño de experimentos en los que el investigador controla o manipula directamente las variables independientes para determinar el efecto sobre las variables dependientes.
MANOVA proporciona las herramientas necesarias para juzgar la fiabilidad de cualquier efecto observado.
En MANOVA, la hipótesis nula contrastada es la igualdad de vectores de medias de variables dependientes múltiples entre los grupos.
En esta técnica el investigador tiene dos combinaciones de variables , una construido a partir de las variables dependientes y otra a partir de las variables independientes.
El aspecto propio de MANOVA es que la combinación lineal optimiza las medidas dependientes múltiples dentro de un valor único que maximiza las diferencias entre los grupos.

Supuestos de MANOVA.
Para que los procedimientos de los contrastes multivariantes de MANOVA sean válidos, se deben cumplir tres supuestos: las observaciones deben ser independientes, las matrices de

varianzas – covarianzas deben ser iguales para todos los grupos de tratamientos y el conjunto de las *p* variables dependientes debe seguir una distribución normal. Además de estos supuestos estadísticos deben considerarse la linealidad y la multicolinealidad de la combinación lineal de las variables, aspectos que afectan los posibles efectos, y la presencia de datos atípicos, ya que el MANOVA es especialmente sensible a la presencia de estos casos.

Estimación del modelo MANOVA.
El MANOVA proporciona varios criterios con los que valorar las diferencias multivariantes entre los grupos. Los cuatro más conocidos son: *la mayor raíz característica de Roy, la lambada de Wilks (estadístico U), la traza de Hotelling y el criterio de Pillai.* Estos criterios valoran las diferencias entre dimensiones de las variables dependientes.
La mayor raíz característica de Roy mide las diferencias solamente sobre la primera raíz canónica (o función discriminante) entre las variables dependientes.
Las otras tres medidas valoran todas las posibles fuentes de diferencia entre los grupos.
El contraste más comúnmente empleado para la significación global del MANOVA es la lambada de Wilks. Este criterio considera todas las raíces características; es decir, compara si los grupos son de algún modo diferentes sin estar afectados por el hecho de que los grupos son de algún modo diferentes sin estar afectados por el hecho de que los grupos difieran en al menos una combinación lineal de las variables dependientes. Aunque la distribución de la lambada de Wilks es compleja, se tienen buenas aproximaciones para contrastar la significación transformándolo en un estadístico *F*.

Validación de los resultados.
Para la técnica de análisis de la varianza MANOVA la replicación es el principal medio de validación. La especificidad de los tratamientos experimentales permite un amplio empleo del mismo experimento en múltiples poblaciones para evaluar la generalidad de los resultados.

4.2.4. Análisis de Correlaciones Canónicas.

El análisis de correlaciones canónicas (ACC) es una técnica estadística multivariante muy útil cuando se tienen múltiples variables dependientes. Está considerado como el modelo general en el que se basan muchos modelos multivariantes, dado que se pueden emplear tanto datos métricos como no métricos para variables tanto dependientes como independientes [Hair , 1999]
De forma general se puede representar por el siguiente modelo:

$$\underset{\text{(métrica, no métrica)}}{Y_1 + Y_2 + Y_3 + \ldots + Y_n} = \underset{\text{(métrica, no métrica)}}{X_1 + X_2 + X_3 + \ldots + X_n}$$

El ACC es una técnica que facilita el estudio de las interrelaciones entre múltiples variables dependientes y múltiples variables independientes, es decir, el ACC predice simultáneamente múltiples variables dependientes a partir de múltiples variables independientes y establece el menor número de restricciones sobre los datos con que se trabaja.

El ACC puede llevarse a cabo con un amplio rango de objetivos, entre los que podemos citar los siguientes:

1. Determinar si dos conjuntos de variables (medidas sobre los mismos objetos) son independientes uno del otro, o por el contrario determinar la magnitud de la relación existente entre ambos conjuntos de variables.
2. Obtener un conjunto de ponderaciones para cada conjunto de variables (dependientes e independientes), para que las combinaciones lineales de cada conjunto estén correlacionadas de forma máxima.
3. Explicar la naturaleza de cualquiera de las relaciones existentes entre los conjuntos de variables dependientes y variables independientes, generalmente midiendo la contribución relativa de cada variable a las funciones canónicas (relaciones) que son extraídas.

Supuestos básicos del ACC.

El ACC está restringido a la identificación de relaciones lineales, pero puede emplear cualquier variable métrica sin que cumpla el supuesto de normalidad (aunque es recomendable que se mida la normalidad de todas las variables.

La homocedasticidad debe ser estudiada, ya que disminuye la correlación entre las variables. Por último la multicolinealidad entre algunos conjuntos de variables distorsiona la capacidad de la técnica para aislar el impacto de cualquier variable , haciendo que la interpretación sea menos fiable.

Estimación del modelo.

El primer paso en el ACC es la obtención de una o más funciones canónicas formadas por un par de combinaciones lineales, una que representa a las variables independientes y la otra representa a las variables dependientes. El número máximo de funciones canónicas que se puede obtener es igual al número de variables que hay en el conjunto de datos menor, ya sea dependiente o independiente. La obtención de las funciones canónicas de forma sucesiva sigue un procedimiento similar al seguido en el análisis factorial sin rotación, pero centrándose en la explicación de las cantidades máximas de relaciones entre los dos conjuntos de variables, en lugar de en un solo conjunto. El resultado es que el primer par de funciones canónicas refleja la mayor intercorrelación, el siguiente par la segunda mayor correlación y así sucesivamente. Cada par de funciones es ortogonal e independiente respecto a todas las otras funciones obtenidas a partir del mismo conjunto de datos.

La validez de la relación entre los pares de funciones se refleja en la correlación canónica. Cuando se eleva al cuadrado, la correlación canónica representa la cantidad de varianza de un valor teórico explicada por la otra función. A esto también se le puede definir como la cantidad de varianza compartida entre las dos funciones canónicas. La correlaciones canónicas al cuadrado se denominan raíces canónicas.

De las funciones obtenidas se interpretan generalmente aquellas cuyo coeficiente de correlación canónica son estadísticamente significativos para un nivel, normalmente 0.05 ó mayor.

No obstante es recomendable que sean empleados tres criterios de manera conjunta para decidir que funciones canónicas interpretar.

1. Nivel de significación estadística de la función (0.05 el más habitual)

1. Nivel de significación estadística de la función (0.05 el más habitual)
2. Magnitud de las relaciones canónicas (la decisión se basa generalmente en la contribución de los resultados para una mejor comprensión del problema que se está estudiando)
3. Medida de la redundancia de la varianza compartida (el índice de redundancia de Stewart-Love [Stewart y Love, 1968] calcula la cantidad de varianza de un conjunto de variables que puede ser explicada por la varianza de otro conjunto. Este índice sirve como una medida de explicación de la varianza similar al cálculo de R^2 empleado en la regresión múltiple. Proporciona una medida resumen de la capacidad del conjunto de las variables independientes (consideradas en conjunto) para explicar la variación de las variables dependientes (consideradas una a una).

Interpretación de la función canónica.
La interpretación comprende el examen de las funciones para determinar la importancia relativa de cada una de las variables originales en las relaciones canónicas.
El enfoque tradicional para interpretar las funciones canónicas comprende el examen del signo y la magnitud de la ponderación canónica asociada a cada variable en su función canónica. Las variables con ponderaciones relativamente mayores contribuyen más a la función y viceversa. Las variables cuyas ponderaciones tienen signos contrarios presentan una relación inversa entre sí, y las variables con ponderaciones del mismo signo representan relación directa.
Un método alternativo en la interpretación de la función canónica es el método de cargas cruzadas canónicas [Dillon, 1984]. Este procedimiento consiste en correlacionar cada una de las variables dependientes originales observadas directamente con la función canónica independiente y viceversa.
Validación del ACC.
El procedimiento más directo para la validación es crear dos submuestras de los datos y llevar a cabo el análisis en cada submuestra de forma separada . Los resultados se comparan para buscar la igualdad de las funciones canónicas, las cargas canónicas y demás aspectos. Si se observar diferencias importantes deberá realizarse una investigación adicional para asegurar que los resultados son representativos de la población..

4.2.5. Análisis Conjunto.

El análisis conjunto es una técnica de análisis multivariante que se utiliza para entender como los encuestados desarrollan preferencias acerca de productos, servicios, ideas (real o hipotética)
Combinando cantidades separadas del valor que proporciona cada atributo. La utilidad que es la base conceptual para medir el valor en el análisis conjunto, es un juicio subjetivo de preferencia única para cada individuo. Abarca todas las características de un producto o servicio, tanto tangible como intangible, y como tal es la medida de preferencia global.
El análisis conjunto es el único entre los métodos multivariantes en el cual el investigador construye primero un conjunto real o hipotético de bienes y servicios combinando niveles escogidos de cada atributo. Estas combinaciones se presentan a los encuestados, que ofrecen solo sus evaluaciones globales. Por tanto, el investigador pide al encuestado que realice una operación muy real, elegir entre un conjunto de productos, servicios o ideas.

Los encuestados no tienen que pronunciarse por la importancia de un atributo concreto o por lo bien que se ajusta el producto a un atributo específico.
Para lograr el éxito con esta técnica debe describirse el producto o servicio en término tanto de susu atributos como de todos los valores relevantes para cada atributo. Se utiliza el término factor para describir un atributo específico u otra característica del producto o servicio. Los valores posibles para cada factor se denominan niveles. Describimos un producto o servicio respecto a su nivel en el conjunto de factores que lo caracterizan.

El modelo de análisis conjunto puede expresarse como:

$$Y_1 = X_1 + X_2 + X_3 + \ldots + X_n$$

(métrica, no métrica) (no métrica)

Este modelo presenta una gran flexibilidad y unicidad ya que utiliza tanto variables dependientes métricas como no métricas, usa variables predictoras categóricas y los supuestos acerca de las relaciones entre las variables dependientes e independientes son bastantes generales.

En el análisis conjunto el diseño experimental en las decisiones del consumidor tiene dos objetivos:

1. Determinar las contribuciones de las variables predictoras y sus niveles en la determinación de las preferencias del consumidor.

2. Establecer un modelo válido de los juicios del consumidor (los modelos básicos nos permiten predecir la aceptación por parte del consumidor de cualquier combinación de atributos, incluso de aquellos no originalmente evaluados por el consumidor.

Para representar el proceso de valoración del encuestado con precisión deben incluirse todos los atributos que potencialmente crean o sustraen utilidad al producto o servicio. Es esencial que se consideren tanto los factores positivos como los negativos, debido fundamentalmente a que centrarse solo en los valores positivos distorsiona seriamente los juicios de los encuestados y los encuestados pueden emplear inconscientemente factores negativos, aunque no se proporcionen en la encuesta y se invalidaría el experimento. Además deben incluirse todos los factores determinantes que mejor diferencian los objetos.

Metodologías básicas del análisis conjunto.
Método tradicional: Se caracteriza por un modelo aditivo simple que contiene nueve factores estimados para cada individuo como máximo.
Método adaptativo conjunto: desarrollado para dar cabida a un gran número de factores (generalmente más de 30) lo cual no sería posible con el modelo tradicional.
Método basado en la elección: Emplea una forma única de presentar los estímulos (en conjunto en lugar de uno a uno), incluye directamente interacciones y debe ser estimado a nivel agregado.

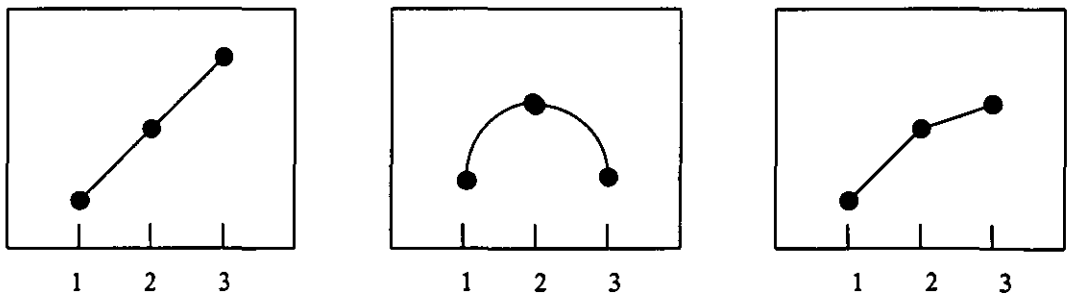
Los fundamentos experimentales del análisis conjunto dan una gran importancia al diseño de los estímulos que van a ser evaluados por lo encuestados . Los factores y niveles deben

ser medidas comunicables y prácticas. Comunicables para propiciar una evaluación realista y práctica, lo que significa que los atributos deben ser distintos y representar un concepto que se puede implementar de forma precisa. No deben ser atributos vagos, ni imprecisos, tales como: bajo, alto o moderado; tales especificaciones se prestan a diferentes percepciones de los individuos con relación a lo que realmente se quiere decir.

Especificación de la forma básica del modelo.
Para que el análisis conjunto explique la estructura de preferencia del encuestado sólo a partir de las evaluaciones conjuntas de un conjunto de estímulos, se deben tomar las siguientes decisiones en relación con el modelo conjunto subyacente.

1. La regla de composición: La regla de composición describe como combina el encuestado los componentes parciales de la utilidad total de los factores para obtener el valor conjunto, las opciones al respecto son las de seleccionar un modelo aditivo frente a uno iterativo.
   Modelo aditivo: En este modelo el encuestado simplemente suma los valores de cada atributo para conseguir el valor total de una combinación de atributos (producto o servicio)
   Modelo de incorporación de los efectos interacción: Esta regla de composición también supone que el consumidor suma los componentes parciales de la utilidad total para todo el conjunto de atributos pero permite que ciertas combinaciones de niveles sean superiores o inferiores a la suma.

2. Selección de las relaciones de los componentes parciales de la utilidad total: El análisis conjunto ofrece tres alternativas, que va desde la más restrictiva (la relación lineal) hasta la menos restrictiva (componentes parciales de la utilidad total separados), con el punto ideal, o modelo cuadrático, entre ambas alternativas. La alternativa de los componentes parciales es la más general, ya que permite estimaciones aisladas para cada nivel.



a) Lineal b) Cuadrática c) Componentes Parciales Separados

Figura 4.4 Tipos de relaciones de los componentes parciales de utilidad. Gráficos de nivel vs. preferencia

Tipo de presentación de los estímulos.
Los métodos de trade–off, perfil completo y comparación pareada son las tres técnicas de presentación de los estímulos más frecuentes en el análisis conjunto.

Método de presentación trade-off: Este método compara dos atributos al mismo tiempo mediante clasificaciones de niveles y utilizando todas las posibles combinaciones de atributos. Tiene la ventaja de ser sencillo y fácil para el encuestado y evita la sobrecarga de información al presentar sólo dos atributos al mismo tiempo.

Método de presentación de perfil completo: Es este el método de presentación más habitual, principalmente por su realismo en la percepción y su capacidad para reducir el número de comparaciones. Cada estímulo se describe por separado. Se obtienen pocos juicios, pero cada uno es más complejo y pueden ser clasificados o calificados.

Método de presentación de combinaciones pareadas: Este método combina los dos anteriores. La combinación pareada es una comparación de dos perfiles (cada uno con múltiples atributos, aunque no necesariamente todos) utilizando a menudo el encuestado una escala de calificación para indicar la fuerza de la preferencia por un perfil sobre otro.

Una de las preocupaciones de cualquier estudio de análisis conjunto es la carga que se pone en el encuestado debido al número de estímulos conjuntos evaluados. Una revisión reciente de los estudios de marketing con análisis conjunto encontró que los encuestados podían completar fácilmente hasta 20 evaluaciones conjuntas. Después de muchas evaluaciones, las respuestas empiezan a ser menos creíbles y menos representativas de la estructura de preferencia subyacente.

Supuestos básicos del análisis conjunto.
El análisis conjunto tiene el grupo de supuestos menos restrictivos del grupo de métodos de dependencia. El diseño experimental estructurado y la naturaleza generalizada del modelo hace innecesarios la mayoría de los tests realizados en otros métodos. Por tanto los tes de normalidad, homocedasticidad e independencia no son necesarios. El uso de estímulos con bases estadísticas asegura que la investigación no está confundida y que los resultados son interpretables bajo la regla de composición asumida.

Especificación de supuestos con relación a los factores.
Número de factores: El número de factores incluidos en el análisis afecta directamente la eficiencia estadística y la fiabilidad de los resultados. A medida que se añaden más factores y niveles, el creciente número de parámetros a estimar exige o bien un número mayor de estímulos o bien una reducción de la fiabilidad de los parámetros.

$$\text{Número mínimo de estímulos} = \text{Número total de niveles para todos los factores} - \text{Número de factores} + 1$$

Multicolinealidad entre factores: Es este un problema que debe ser solucionado. La correlación entre los factores denota una falta de independencia conceptual entre ellos. En tales casos, los parámetros estimados se ven afectados igual que en la regresión.

El papel único que representa el precio como factor: El precio tiene un alto grado de correlación inter-atributo con otros factores, por lo que debe intentarse prevenir los impactos y ajustar los diseños y la interpretación exigida.

Especificación de supuestos en relación con los niveles.
Número equilibrado de niveles: Debe tratarse siempre de equilibrar o igualar el número de niveles para todos los factores. Se ha encontrado que la importancia relativa estimada de una variable aumenta a medida que el número de niveles lo hace., incluso si los extremos siguen siendo los mismos.

Rango de los niveles de un factor: El rango de los niveles debe fijarse fuera de los valores existentes pero no en un nivel improbable. Con eso se reduce la correlación inter-atributo , pero también se puede reducir la credibilidad , por lo que los niveles no deben ser muy extremos.

Interpretación de los resultados:
La forma normal para interpretar el análisis conjunto es la desagregada, esto implica, modelizar a cada encuestado separadamente y los resultados del modelo se examinan para cada encuestado. El método más común de interpretación es un examen de las estimaciones de los componentes parciales para cada factor, evaluando su magnitud y su pauta tanto a efectos de relevancia práctica como a efecto de correspondencia con relaciones teóricas entre niveles. Cuanto mayor sea el componente parcial (positivo o negativo), mayor será el impacto que tenga sobre la utilidad total. Los valores de los componentes parciales de la utilidad total pueden ser representados gráficamente para identificar las pautas.

Validación de los resultados.
Los resultados del análisis conjunto pueden validarse tanto internamente cono externamente. La validación interna implica la confirmación de que la regla de composición seleccionada (aditiva frente a interactiva) es la apropiada. La validación externa implica en general la capacidad del análisis conjunto de predecir elecciones efectivas, y de forma más específica la representatividad de la muestra.

4.2.6 Modelo de Ecuaciones Estructurales.

Producto de una evolución de la modelización multiecuacional desarrollada principalmente en la econometría y fusionada con los principios de medición de la psicología y la sociología, el modelo de ecuaciones estructurales (SEM) se ha convertido en una herramienta integral tanto en la investigación académica como en la práctica.
El SEM examina simultáneamente una serie de relaciones de dependencia. Este conjunto de relaciones cada una con variables dependientes e independientes, es la base de del SEM y se puede formular básicamente de la siguiente forma:

$$Y_1 = X_{11} + X_{12} + X_{13} + \ldots + X_{1n}$$
$$Y_2 = X_{21} + X_{22} + X_{23} + \ldots + X_{2n}$$
$$\vdots \quad \vdots \quad \vdots \quad \vdots$$
$$Y_m = X_{m1} + X_{m2} + X_{m3} + \ldots + X_{mn}$$

(métrica) (métrica, no métrica)

Todas las técnicas SEM se distinguen por dos características: estimación de relaciones de dependencia múltiple y cruzada SEM estima una serie de ecuaciones de regresión múltiples distintas pero interrelacionadas mediante la especificación del modelo estructural utilizado por el programa estadístico), y la capacidad de representar conceptos no observados en estas relaciones (variables latentes) y tener en cuenta el error de medida en el proceso de estimación.
El resultado final del SEM es siempre la evaluación de una serie de relaciones. Sin embargo, esto se puede conseguir de muchas maneras. En nuestro caso definiremos tres estrategias de aplicación del SEM: de modelización confirmatoria, de modelos rivales y de desarrollo del modelo.

Estrategia de Modelización Confirmatoria: Es la aplicación, más directa de los modelos de ecuaciones estructurales, donde el investigador especifica un modelo aislado y SEM se utiliza para evaluar su significación estadística.

Estrategia de Modelos Rivales: Como medio de evaluar el modelo estimado con modelos alternativos, las comparaciones de modelos conjuntos pueden llevarse a cabo en una estrategia de modelos rivales. El test más fuerte de un modelo propuesto es identificar y contrastar los modelos rivales que representan las verdaderas y distintas relaciones hipotéticas estructurales.

Estrategia de desarrollo del modelo: Esta estrategia difiere de la anteriores en que aunque se propone un modelo, el propósito del esfuerzo de modelización es mejorarlo a través de modificaciones de los modelos de medida y/o estructurales.

La modelización de ecuaciones estructurales se basa en relaciones causales, en la que el cambio en una variable se supone que produce un cambio en otra variable. La fuerza y convicción con que el investigador puede asumir la causación entre dos variables no descansa en los métodos analíticos escogidos sino en la justificación teórica ofrecida por los análisis.

Los diagramas de secuencias son especialmente útiles en el diseño de una serie de relaciones causales. Este tipo diagrama permite presentar no sólo relaciones predictivas entre constructos (entre variables dependiente-independiente), sino también relaciones asociativas (correlaciones) entre los constructos e incluso entre los indicadores.

En la figura 4.5 se muestra un ejemplo de relaciones representadas por diagramas de secuencias. Se presentan tres constructos dependientes cada uno relacionado con el resto y

con los constructos independientes. La estimación de todas las ecuaciones simultáneamente sólo es posible a través del SEM.

Relaciones causales

$X_1\ X_2 \rightarrow Y_1$

$X_2\ X_3\ Y_1\ Y_3 \rightarrow Y_2$

$Y_1\ Y_2 \rightarrow Y_3$

v. independientes v. dependientes

Diagrama de relaciones

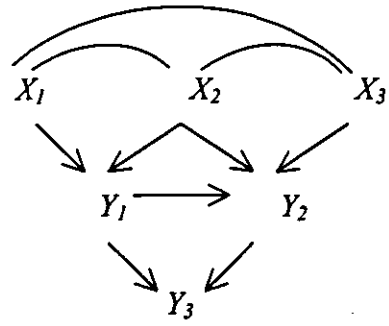


Figura 4.5 Ejemplo de diagrama de relaciones

Después de desarrollar el modelo teórico y de representarlo en un diagrama de secuencias, se puede pasar a especificar el modelo de manera formal. Esto se hace a través de una serie de ecuaciones que definen las ecuaciones estructurales que vinculan los constructos, el modelo de medida que especifica qué variables miden qué constructo y una serie de matrices que indican cualquier correlación supuesta entre constructos o variables.
Para cada ecuación se considera además un término de error que representa la suma de los efectos debidos a un error de especificación y errores aleatorios de medida.
Basándonos en el ejemplo de la figura 4.5 podemos ejemplificar el proceso de transición del diagrama de secuencias a ecuaciones estructurales.

$$Y_1 = b_1X_1 + b_2X_2 + \varepsilon_1$$
$$Y_2 = b_3X_2 + b_4X_3 + b_5Y_1 + b_6Y_3 + \varepsilon_2$$
$$Y_3 = b_7Y_1 + b_8Y_2 + \varepsilon_3$$

SEM difiere de otras técnicas multivariantes en que sólo utiliza la matriz de varianza covarianza o de correlación como sus datos de entrada. El interés no está en las observaciones individuales sino en el patrón de relaciones entre los encuestados.

Supuestos en el SEM.
Los supuestos básicos de SEM son: observaciones independientes, muestras aleatorias de los encuestados y la linealidad de todas las relaciones, además SEM es muy sensible al incumplimiento de la normalidad multivariante o a una fuerte curtosis de los datos.

Estimación del modelo.
Los intentos iniciales de estimación del modelo de ecuaciones estructurales se realizaron con la regresión de los mínimos cuadrados ordinarios , pero este fue superado rápidamente por la estimación de máximo verosímil (muy utilizada en la mayoría de los programas informáticos), que es eficiente y no sesgada cuando se cumplen los supuestos de normalidad multivariante. Otros métodos desarrollados para hacer frente a la no normalidad han sido: mínimos cuadrados ponderados, mínimos cuadrados generalizados,

asintóticamente libre de distribución; esta última ha recibido recientemente una atención particular debido a su alta insensibilidad a la no normalidad de los datos. Su principal exigencia es un aumento del tamaño de muestra.

Además de la técnica de estimación empleada , se puede también escoger entre varios procesos de estimación, que van desde la estimación directa del modelo, tal y como hemos visto en las restantes técnicas multivariantes , a métodos que generan miles de estimaciones del modelo para las cuales se obtienen los resultados finales del modelo.

Interpretación del modelo.
Una vez obtenido un modelo aceptable, deben examinarse los resultados y su correspondencia con la teoría propuesta. ¿Están corroboradas y son estadísticamente significativas las principales relaciones de la teoría? ¿ Añaden los modelos rivales mayor perspectiva sobre las formulaciones alternativas de la teoría como para que puedan ser tenidas en cuenta?¿Están todas las relaciones en la dirección supuesta? Todas estas cuestiones pueden ser contestadas a partir de los resultados empíricos.

CAPÍTULO 5.

Aplicación de la Informática en el Análisis Multivariante.

El análisis multivariante tiene sus orígenes en los trabajos de los matemáticos del siglo pasado que desarrollaron el álgebra linear y la geometría multidimensional que forman la base de muchos de los métodos multivariantes pero sólo con el desarrollo de la computación en los últimos 20 años se hizo posible el desarrollo y aplicación de estas técnicas por especialistas e investigadores de muchas disciplinas, por lo que se hace prácticamente imposible analizar la aplicación de las técnicas multivariantes sin referirnos a algunos de los principales paquetes de programas estadísticos diseñados para computadoras con este propósito. Este precisamente constituye el objetivo de este capítulo.

5.1. Principales software para aplicación de las técnicas de análisis multivariante.

Tal y como hemos señalado anteriormente el análisis multivariante se llevan a cabo a través de programas diseñados para computadoras, utilizando alguno de los paquetes estadísticos que con este fin se han desarrollado. Cada paquete tiene su propio formato, instrucciones, procedimientos y características.

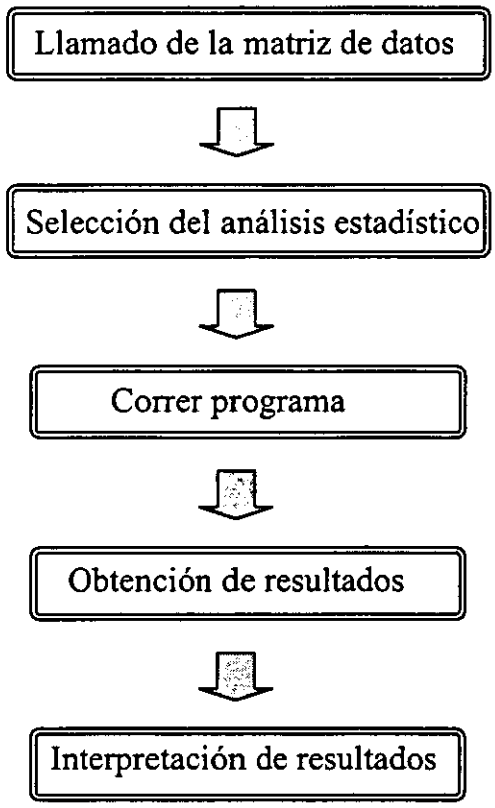
Entre los paquetes estadísticos más conocidos hoy en día se encuentran el SPSS (*Statistical Package for the Social Sciences*), desarrollado en la Universidad de Chicago, al cual nos referiremos más adelante; el SAS (*Statistical Analysis System*), desarrollado en la Universidad Estatal de Carolina del Norte y distribuido por SAS Institute, Inc. es muy poderoso y su uso se ha incrementado notablemente; y el BMDP (*Biometrical Computer Programs*), desarrollado por la Universidad de California, Los Ángeles, aunque diseñado por el área biomédica, contiene una gran cantidad de análisis aplicables en otras ciencias.

El SPAD es otro paquete estadístico diseñado específicamente para el análisis multivariante, con amplias posibilidades .

Estos paquetes de programas incluyen una gran cantidad de técnicas estadísticas, algunas incluidas en nuestro objeto de estudio.

Varios paquetes han sido desarrollados para su aplicación específica en el área del análisis multivariante. Ejemplo de estos son los siguientes: CUSTAN, dirigido al análisis cluster, MULTISCAL y MDS(X) para el análisis multidimensional y el LISREL VII enfocado a los modelos de ecuaciones estructurales.

Aunque la variedad de paquetes estadísticos se ha ido incrementando en los últimos años y cada uno presenta su forma específica de trabajo es posible establecer un procedimiento general para el análisis de datos con muchos de los paquetes existentes y que puede resumirse de la forma siguiente:



Principales aplicaciones del análisis multivariante posibles con los software estadísticos más comercializados.

SPSS par Windows.

- *Análisis Factorial*
- *Análisis de Componentes Principales*
- *Regresión Múltiple*
- *Análisis Discriminante*
- *Regresión Logística*
- *MANOVA*
- *Análisis Conjunto*
- *Correlación Canónica*
- *Análisis de grupos*
- *Análisis multidimensional*
- *Análisis de Correspondencias*

SAS

- *Análisis de Componentes Principales*
- *Análisis factorial Común*

- *Regresión Múltiple*
- *Análisis Discriminante*
- *Análisis de Regresión Logística*
- *MANOVA*
- *Correlaciones Canónicas*
- *Análisis de Grupos*

LISREL VIII

- *Análisis Factorial*
- *Modelo de Ecuaciones Estructurales*

PC – MDS y MULTISCAL

- *Análisis Multidimensiona*
- *Análisis de Correspondencia*

SPAD

- *Análisis Factorial*
- *Análisis de Componentes Principales*
- *Análisis de Correspondencia*
- *Análisis de Grupos (Clasificación a partir de un análisis factorial previo)*

CLUSTAN

- *Análisis de Grupos*

5.2. **Paquete Estadístico** *SPSS.*

Referencia de Manual.
Visauta Vinacua, Bienvenido (1998) Análisis Estadístico con SPSS para Windows. Vol II. Estadística Multivariante, McGraw Hill.

SPSS en su versión para Windows trabaja de manera muy amigable. El usuario puede seleccionar las opciones más apropiadas para su análisis, de forma similar a como se hace en otros programas que se encuentran en este ambiente.
El procedimiento de trabajo de forma general es el siguiente.

1. Selección de la ventana de SPSS.
2. SPSS nos muestra en pantalla un formato para matriz de datos y un conjunto de opciones desplegables, tal y como se muestra en la siguiente figura.

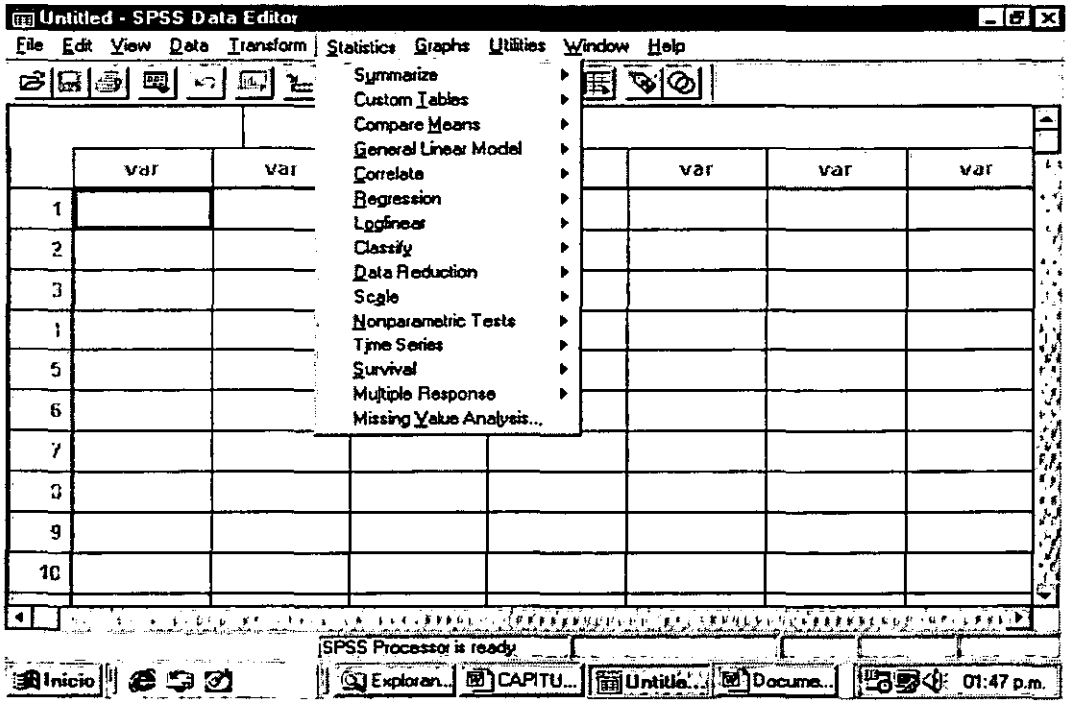


Figura 5.1 Posibilidades de la opción "statistics".

3. Definimos cada variable (seleccionando "data editor",y luego "define variate")
4. Llenado de la matriz de datos
5. Seleccionar la opción de "Statistics" y elegir la prueba apropiada. Las posibilidades que nos brinda esta opción son las siguientes:

- *Summarize (Sumarios)*
- *Custom Tables*
- *Compare Means (Comparar Medias)*
- *General Linear Model → (MANOVA)*
- *Correlate (Correlación)*
- *Regression (Regresión)*
- *Loglinear →(Análisis Loglineal)*
- *Classify (Clasificación)*
  - *Discriminant (Análisis Discriminante)*
  - *Hierarchical Cluster (Cluster Jerárquico)*
  - *K-Means Cluster (Cluster no Jerárquico)*

- *Data reduction Reducción de datos*
  Factor (Análisis factorial, Análisis de Componentes Principales)
  Correspondence Análisis (Análisis de Correspondencia)
  Optimal Scaling (Escalamiento Óptimo) →(Correlación Canónica)
- *Scale (Escala)*
  Multidimensional Scaling (Análisis Multidimensional)
- *Nonparametric test (Pruebas no Paramétricas)*
- *Time Series (Series de Tiempo)*
- *Survival*
- *Multiple Response (Respuestas Múltiples)*
- *Missing Value Analysis... (Análisis de valores ausentes)*

El cálculos es efectuado por la computadora. Los resultados se muestran en pantalla y pueden mandarse a imprimir.

Ejemplos de aplicación de este programas se muestran en el capítulo dedicado a los casos de aplicación.

CAPÍTULO 6.

Ejemplos de Aplicación del Análisis Multivariante.

En este , el capítulo final del presente trabajo, se recogen dos ejemplos ilustrativos de aplicación de las técnicas multivariantes para el análisis de datos. Se incluye un ejemplo de las técnicas descriptivas de interdependencia (técnica factorial), un ejemplo de las técnicas explicativas de dependencia (análisis discriminante) con lo cual se persigue el objetivo de mostrar la aplicación práctica de los aspectos tratados en los capítulos precedentes, se incluyen además los aspectos fundamentales del manejo del paquete *SPSS* en la obtención de los principales resultados.

6.1 CASO 1. Aplicación del Análisis de Componentes Principales.

Definición del Problema y los Objetivos.
Se examinan los resultados de una encuesta realizadas sobre siete atributos o características de una empresa para comprender si estas pueden ser agrupadas y reducir las siete variables a un número menor [Hair , 1999]. Agrupando las percepciones , la empresa dispondrá de un panorama que le permitirá comprender a sus clientes y lo que ellos piensan sobre la empresa. Si las siete variables pueden representarse en un número menor de variables compuestas, se facilita cualquier análisis posterior con otra técnica.

Los datos de partida están constituidos por 100 observaciones de las siete variables referidas a la empresa que se analiza. Los encuestados dieron una puntuación sobre cada atributo (ver base de datos en el anexo 1). La descripción de cada variable se da a continuación.

Etiqueta	Denominación	Descripción	Tipo
X_1	Velocidad de Entrega	Tiempo que transcurre hasta que se entrega el producto, una vez que se ha confirmado el pedido	Métrica
x2	Nivel de Precio	Nivel de precio percibido por los clientes industriales.	Métrica
X_3	Flexibilidad de Precio	Disposición en los representantes de la empresa para negociar el precio de todas las compras.	Métrica
X_4	Imagen de Fabricante	Imagen conjunta fabricante - distribuidor	Métrica
X_5	Servicio Conjunto	Nivel conjunto de servicio necesario para mantener una relación satisfactoria entre el suministrador y el comprador	Métrica
X_6	Imagen de Fuerza de Venta	Imagen conjunta de la fuerza de ventas del fabricante.	Métrica
X_7	Calidad del Producto	Nivel de calidad percibido en un producto particular	Métrica

Tabla 6.1 Descripción de las variables de la base de datos.

Para realizar el procesamiento de los datos se seleccionó el paquete estadístico SPSS 8.0. Se comienza editando y cargando el archivo de datos y entrando en ***Statistics/Data Reduction/Factor.*** Se obtiene el cuadro de salida (pantalla) *Factor Análisis* (figura 6.1), se seleccionan las variables que intervendrán en el análisis , en nuestro caso todas, y se eligen las diferentes opciones que diseñan el análisis en cada uno de los subcuadros (*Descriptives, Extraction, Rotation, Score, Options*), y que oportunamente serán explicadas.
Pulsando *OK* en el cuadro de dialogo principal se ejecuta el procedimiento, obteniéndose las salidas que a continuación se analizan.

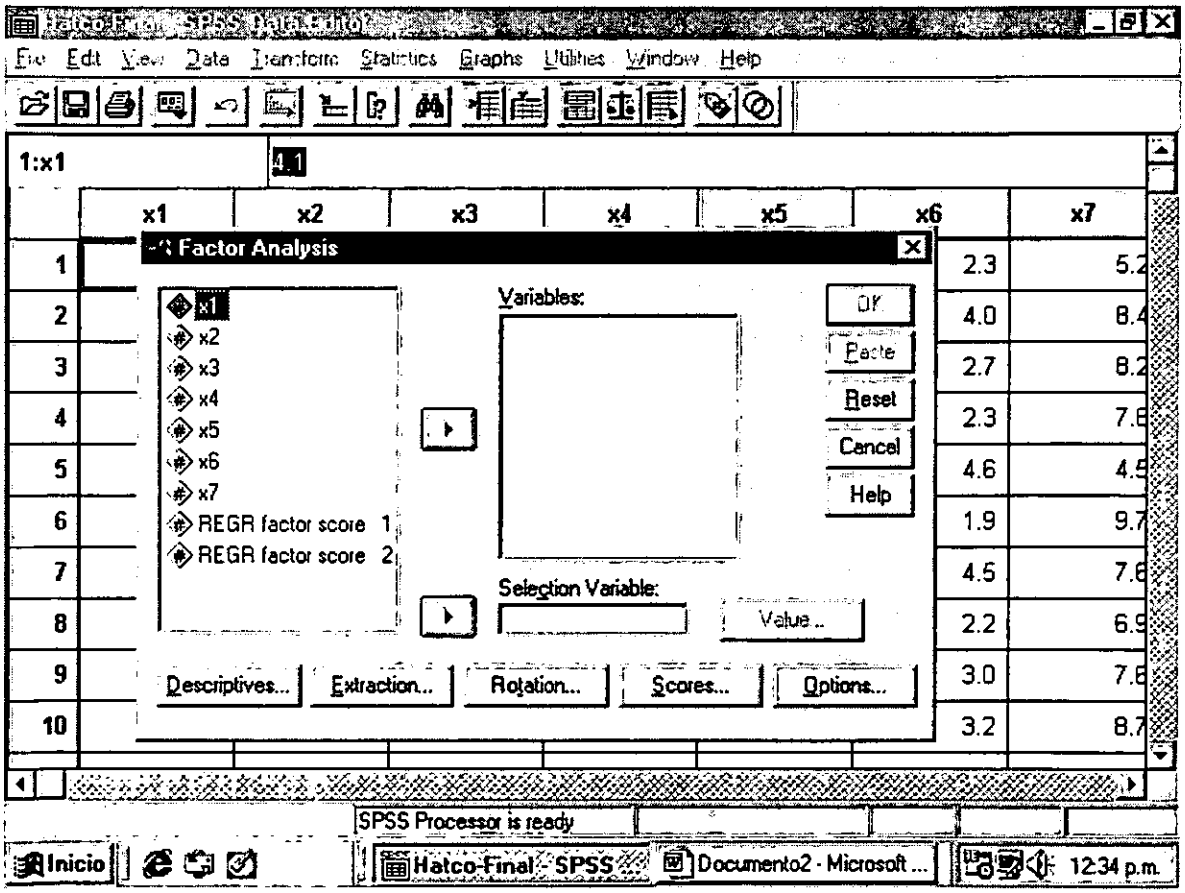


Figura 6.1 Cuadro de diálogo principal del análisis factorial. Programa SPSS versión 8.0

Factor Analysis

En la tabla *Descriptive Statistics* aparecen la mediana y la desviación estándar de las siete variables considerando los cien sujetos de la muestra. Para obtener esta tabla de salida en el subcuadro *Descriptive* seleccionamos la opción *Univariante Descriptive.*

Descriptive Statistics

	Mea	Std. Deviation	Analysis N
X1	3.51	1.321	100
X2	2.36	1.196	100
X3	7.89	1.387	100
X4	5.24	1.131	100
X5	2.91	.751	100
X6	2.66	.771	100
X7	6.97	1.585	100

A continuación aparece la matriz de correlación (Correlation Matriz)entre variables y el grado de significación de estos coeficientes en un contraste univariante. Al pie de la tabla se da el determinante de la matriz de correlación.

Correlation Matrix

		X1	X2	X	X4	X5	X	X7
Correlation	X1	1.000	-.349	.50	.050	.612	.07	-.483
	X2	-.349	1.000	-.48	.272	.513	.18	.470
	X3	.509	-.487	1.00	-.116	.067	-.03	-.448
	X4	.050	.272	-.11	1.000	.299	.78	.200
	X5	.612	.513	.06	.299	1.000	.24	-.055
	X6	.077	.186	-.03	.788	.241	1.00	.177
	X7	-.483	.470	-.44	.200	-.055	.17	1.000
Sig. (1-tailed)	X1		.000	.00	.309	.000	.22	.000
	X2	.000		.00	.003	.000	.03	.000
	X3	.000	.000		.125	.255	.36	.000
	X4	.309	.003	.12		.001	.00	.023
	X5	.000	.000	.25	.001		.00	.293
	X6	.223	.032	.36	.000	.008		.039
	X7	.000	.000	.00	.023	.293	.03	

a Determinant = 2.679E-03

Análisis de los supuestos.
Como ya conocemos los supuestos estadísticos afectan al análisis factorial en la medida en que afectan a las correlaciones obtenidas. Incumplimientos en la normalidad, la homocedasticidad y la linealidad pueden reducir las correlaciones entre variables. Es importante que todas las variables tengan al menos un coeficiente de correlación significativo en la matriz.
De la inspección visual de la matriz de correlaciones resulta que 12 de las 21 correlaciones son significativas al nivel de 0.01 para un 57%.
La siguiente información de salida del programa muestra la inversa de la matriz de correlaciones, los KOM (Kaiser-Meyer-Olkin) y el test de esfericidad de Bartlett.

Inverse of Correlation Matrix

	X	X2	X3	X	X5	X6	X
X1	35.74	32.158	.140	1.50	-38.694	-.590	-.12
X2	32.15	31.597	1.118	1.27	-36.298	-.413	-1.00
X3	.14	1.118	1.645	.20	-.775	-.179	.22
X4	1.50	1.277	.207	2.87	-1.942	-2.134	-.08
X5	-38.69	-36.298	-.775	-1.94	43.834	.562	.73
X6	-.59	-.413	-.179	-2.13	.562	2.697	-.19
X7	-.12	-1.005	.227	-.08	.735	-.191	1.60

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.446
Bartlett's Test of Sphericity	Approx. Chi-Square	567.541
	df	21
	Sig.	.000

El test de Bartlett se utiliza para verificar si la matriz de correlación es una matriz identidad, es decir, si todos los coeficientes de la diagonal son iguales a la unidad y los extremos iguales a cero. Este estadístico se obtiene a partir de la transformación χ^2 del determinante de la matriz de correlación y permite valorar la significación de dicha matriz.
En nuestro caso las correlaciones cuando se toman conjuntamente, son significativas

Contraste de esfericidad de Bartlett: 567,541
Significancia: 0.000

El índice KOM nos compara los coeficientes de correlación de Pearson con los coeficientes de correlación parcial entre variables.
Si el KOM se encuentra próximo a la unidad el análisis factorial es un procedimiento adecuado. En cambio, valores pequeños en este indicador nos dan a entender todo lo contrario.
Este contraste se puede valorar con igual criterio que la medida de adecuación muestral MSA (Measures of Sampling Adequacy) por lo que su valor de 0.446, inferior a 0.50, cae en un rango no aceptable. Por otro lado las MSA para cada variable (diagonal de la matriz anti-imagen de correlación) muestra que las variables etiquetadas como X_1, X_2 Y X_5 tienen valores no aceptables por debajo de 0.50 (0.344, 0.330 Y 0.288 respectivamente).
Con el propósito de elevar estos valores y lograr la adecuación para el análisis factorial, se procederá a eliminar la variable X_5 del análisis y ejecutar nuevamente el procedimiento para el conjunto restante de variables.
Ejecutando el procedimiento desde el inicio para el nuevo conjunto de variables (X_1, X_2, X_3, X_4, X_6, X_7) obtenemos los siguientes resultados.

Factor Analysis

Correlation Matrix

		X1	X2	X	X4	X6	X
Correlation	X1	1.000	-.349	.50	.050	.077	-.48
	X2	-.349	1.000	-.48	.272	.186	.47
	X3	.509	-.487	1.00	-.116	-.034	-.44
	X4	.050	.272	-.11	1.000	.788	.20
	X6	.077	.186	-.03	.788	1.000	.17
	X7	-.483	.470	-.44	.200	.177	1.00
Sig. (1-tailed)	X1		.000	.00	.309	.223	.00
	X2	.000		.00	.003	.032	.00
	X3	.000	.000		.125	.367	.00
	X4	.309	.003	.12		.000	.02
	X6	.223	.032	.36	.000		.03
	X7	.000	.000	.00	.023	.039	

a Determinant = .117

Inverse of Correlation Matrix

	X	X2	X3	X	X6	X7
X1	1.58	.116	-.544	-.20	-.093	.527
X2	.11	1.539	.476	-.33	.052	-.397
X3	-.54	.476	1.631	.17	-.170	.240
X4	-.20	-.331	.173	2.79	-2.109	-.051
X6	-.09	.052	-.170	-2.10	2.690	-.201
X7	.52	-.397	.240	-.05	-.201	1.594

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.665
Bartlett's Test of Sphericity	Approx. Chi-Square	205.965
	df	15
	Sig.	.000

Anti-image Matrices

		X1	X2	X	X4	X6	X
Anti-image Covariance	X	.629	4.752E-02	-.21	-4.655E-02	-2.184E-02	.20
	X	4.752E-02	.650	.19	-7.706E-02	1.260E-02	-.16
	X	-.210	.190	.61	3.791E-02	-3.864E-02	9.232E-0
	X	-4.655E-02	-7.706E-02	3.791E-0	.358	-.281	-1.154E-0
	X	-2.184E-02	1.260E-02	-3.864E-0	-.281	.372	-4.680E-0
	X	.208	-.162	9.232E-0	-1.154E-02	-4.680E-02	.62
Anti-image Correlation	X	.721	7.433E-02	-.33	-9.808E-02	-4.515E-02	.33
	X	7.433E-02	.787	.30	-.160	2.565E-02	-.25
	X	-.338	.301	.74	8.092E-02	-8.093E-02	.14
	X	-9.808E-02	-.160	8.092E-0	.542	-.769	-2.434E-0
	X	-4.515E-02	2.565E-02	-8.093E-0	-.769	.532	-9.689E-0
	X	.331	-.253	.14	-2.434E-02	-9.689E-02	.77

a Measures of Sampling Adequacy(MSA)

Siguiendo un razonamiento similar se obtienen las siguientes conclusiones parciales:

- Coeficientes de Correlación: 8 de 15 correlaciones significativas
- El índices KOM = 0.665 aceptable
- El test de Bartlett con un $\chi^2 = 205.965$ y un nivel de significancia de 0.0001
- MSA adecuados para todas las variables
- Todas las correlaciones parciales (excepto X_4X_6) bastante bajas

Todo lo anterior nos lleva a la conclusión parcial de que el análisis factorial que sigue a continuación resulta a priori pertinente y puede proporcionarnos resultados satisfactorios.

Extracción de factores.

Como se ha planteado la finalidad del análisis factorial es la de poder llegar a interpretar una matriz de correlaciones. Esta matriz se transforma para obtener la matriz de factores. Las cargas de cada variable sobre cada factor se interpretan para identificar la estructura de las variables.

El programa toma por defecto el método de extracción de factores de componentes principales que es precisamente el que vamos a utilizar (subcuadro de dialogo *Extraction*).

De entrada, la decisión respecto al número de factores que deseamos para representar los datos puede adoptarse desde una doble vía que es la que aparece en el subcuadro de diálogo *Extraction* opción *Extract*. Por defecto el sistema extraerá tantos factores como haya en la solución inicial con valores propios o autovalores superiores a 1 (criterio de la raíz latente). Evidentemente, podemos cambiar el valor por defecto correspondiente a *engenvalue over.*

La segunda posibilidad corresponde al botón *Number of Factors* y consiste sencillamente en fijar un número entero determinado de factores, siempre inferior al número de variables.

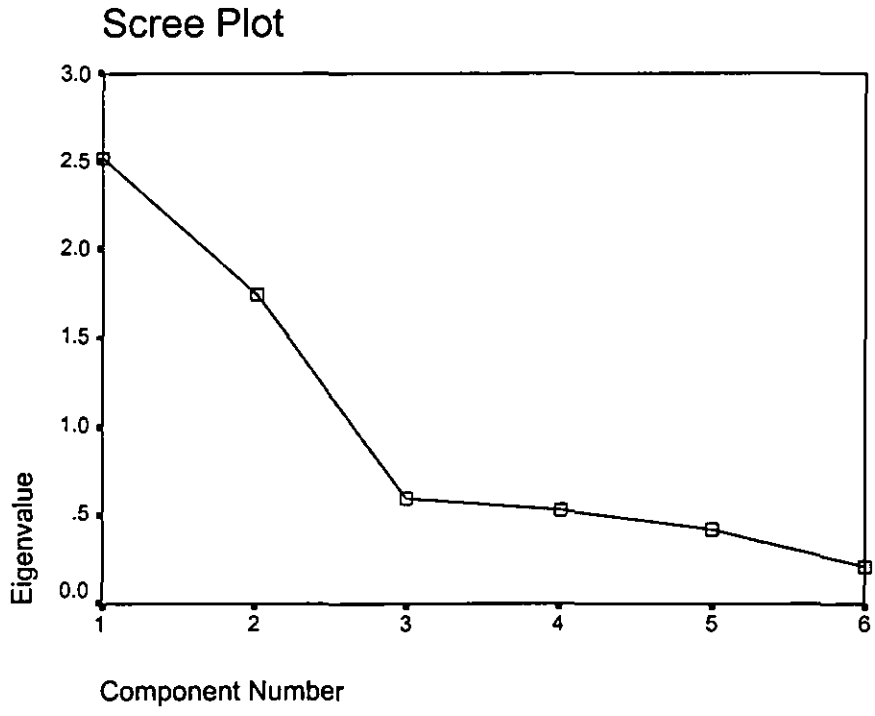
La tabla de salida Total Variance Explained recoge en porcentajes individuales y acumulados la proporción de varianza total explicada por cada factor, tanto para la solución no rotada como para la rotada.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.513	41.892	41.892	2.513	41.892	41.892	2.370	39.497	39.497
2	1.740	28.992	70.883	1.740	28.992	70.883	1.883	31.386	70.883
3	.597	9.958	80.842						
4	.530	8.826	89.668						
5	.416	6.929	96.596						
6	.204	3.404	100.000						

Extraction Method: Principal Component Analysis.

Aplicando el criterio de la raíz latente se incluyen dos componentes en el modelo que explican exactamente el 70.883% de la variabilidad total de las seis variables, lo que puede considerarse un buen porcentaje.
Los resultados muestran además un gráfico de los autovalores contra número total de factores



La tabla Component Matrix muestra los coeficientes utilizados para expresar cada variable estandarizada en términos de los dos factores de modelo (cargas factoriales sobre cada variable para cada factor).

Component Matrix

	Component	
	1	2
X7	.767	-.168
X2	.759	-6.790E-02
X3	-.730	.337
X1	-.627	.514
X6	.425	.832
X4	.494	.798

Extraction Method: Principal Component Analysis.
a 2 components extracted.

Los factores con pesos factoriales más elevados en términos absolutos indican una relación estrecha con las variables. El ideal desde el punto de vista del análisis factorial es encontrar un modelo en el que todas las variables saturen en algún factor, es decir, pesos factoriales altos en uno y bajos en el resto.
El la tabla de referencia aparecen las variables ordenadas de mayor a menor peso ó carga factorial, comenzando por el primer factor (se corresponde con la opción *Sorted by Size* seleccionada en el subcuadro de dialogo *Options*)

Para determinar en que medida dos factores son capaces de explicar las variables originales, podemos considerar la comunalidad que aparece en la diagonal de la tabla Reproduced Correlations

Reproduced Correlations

		X1	X2	X	X4	X6	X
Reproduced Correlation	X1	.658	-.511	.63	.101	.161	-.56
	X2	-.511	.580	-.57	.321	.266	.59
	X3	.631	-.576	.64	-9.188E-02	-3.026E-02	-.61
	X4	.101	.321	-9.188E-0	.882	.874	.24
	X6	.161	.266	-3.026E-0	.874	.872	.18
	X7	-.567	.593	-.61	.245	.187	.61
Residual	X1		.161	-.12	-5.042E-02	-8.417E-02	8.410E-0
	X2	.161		8.920E-0	-4.854E-02	-7.982E-02	-.12
	X3	-.121	8.920E-02		-2.422E-02	-4.058E-03	.16
	X4	-5.042E-02	-4.854E-02	-2.422E-0		-8.577E-02	-4.506E-0
	X6	-8.417E-02	-7.982E-02	-4.058E-0	-8.577E-02		-9.210E-0
	X7	8.410E-02	-.123	.16	-4.506E-02	-9.210E-03	

Extraction Method: Principal Component Analysis.
a Residuals are computed between observed and reproduced correlations. There are 10 (66.0%) nonredundant residuals with absolute values > 0.05.
b Reproduced communalities

Por ejemplo la comunalidad de $X_3 = 0.646$ indica que tiene menos en común con las otras variables incluidas en el análisis de lo que lo hace la variable X_4 con comunalidad de 0.88.
El primer factor da cuenta de la mayor cantidad de varianza y es un factor general, en el que todas las variables tienen cargas altas. Las cargas del segundo factor muestran tres variables (X_1, X_4, X_6) que también tienen cargas altas. Dado este patrón de altas cargas factoriales, la interpretación resulta difícil por lo que se procede a rotar la matriz factorial.

De la rotación debe resultar un patrón factorial más sencillo y más significativo.
En el cuadro de diálogo *Rotation* existen varios procedimientos: VARIMAX, EQUAMAX y QUARTIMAX.
El método seleccionado fue el VARIMAX y los resultados aparecen en la tabla Rotated Component Matriz.

Rotated Component Matrix

	Component	
	1	2
X3	-.804	-1.058E-02
X1	-.787	.194
X7	.764	.179
X2	.714	.266
X6	2.537E-02	.934
X4	.102	.933

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.
a Rotation converged in 3 iterations.

La cantidad de varianza total es la misma en la solución rotada pero el patrón de cargas factoriales y el porcentaje de varianza para cada factor es diferente.
La interpretación de la matriz factorial se ha simplificado, en la solución rotada las variables X_1, X_2, X_3 y X_7 cargan significativamente sobre el factor 1 y las variables X_4 y X_6 sobre el factor 2. Ninguna variable carga significativamente sobre los dos factores
El punto de corte de las cargas a efectos interpretativos es 55% (ver tabla 4.1 Criterios para seleccionar cargas factoriales significativas) de tal forma la interpretación nos da que el primer factor tiene 4 cargas significativas y el segundo tiene 2.

Factor 1 (Valor Básico)
X_2 Nivel de Precio (+ 0.714)
X_7 Calidad del Producto (+ 0.764)
X_1 Rapidez en el Envió (– 0.787)
X_3 Flexibilidad de Precios (– 0.804)
Este factor representa una concesión entre las percepciones del precio o calidad del producto y las percepciones de rapidez en el envío y flexibilidad de precio.

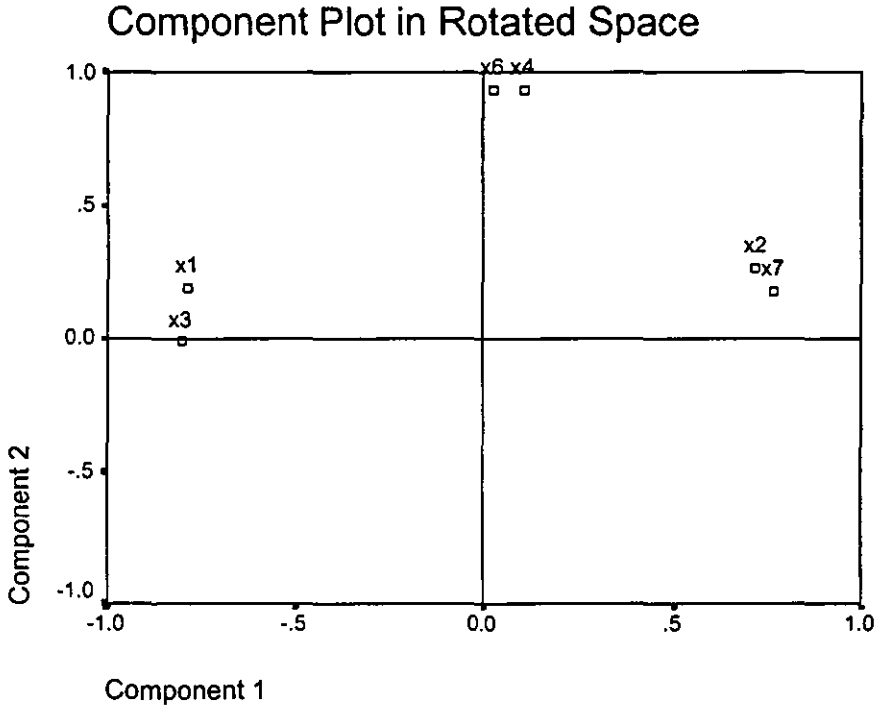
Factor 2 (Imagen)
X_4 Imagen del Productor (+ 0.933)
X_6 Imagen de los Vendedores (+ 0.934)
En este factor ambas variables se relacionan con componentes de imagen, ambas variables tienen el mismo signo, por lo que suponen percepciones bastantes similares.

El gráfico Component Plot in Rotated Space muestra esta solución rotada VARIMAX (opción *Loading Plot(s)* del subcuadro de *Rotation*).

Component Transformation Matrix

Componen t		2
1	.90	.431
2	-.43	.902

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.



Puntuación Factorial.

Puesto que la finalidad final del análisis factorial es reducir el número de variables a un número pequeño de nuevos factores, es aconsejable estimar las puntuaciones factoriales de cada sujeto.

En el subcuadro de dialogo *Factor Score* aparecen las diversas técnicas que ofrece el programa para obtener los coeficientes de las puntuaciones factoriales. Seleccionando el método *Regression* y las opciones *Save as Variables* y *Display Factor Score Coefficient Matriz,* con componentes principales y sin rotación, obtenemos los coeficientes de la tabla Component Score Coefficient Matrix (matriz de coeficientes de puntuaciones factoriales)

Component Score Coefficient Matrix

	Component	
	1	2
X1	-.352	.159
X2	.289	.095
X3	-.345	.050
X4	-.020	.499
X6	-.053	.504
X7	.317	.044

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. Component Scores.

Component Score Covariance Matrix

Component	1	2
1	1.000	.000
2	.000	1.000

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. Component Scores.

A su vez en la base de datos habremos generado dos columnas correspondiente a las puntuaciones factoriales de cada sujeto en cada uno de los dos factores del modelo que podrían sustituir a las seis variables originales en las aplicaciones de otras técnicas.

Validación.
En este ejemplo dividiendo la muestra en dos y reestimando los modelos para contrastar la comparabilidad, se obtiene que las dos rotaciones VARIMAX son bastante comparables tanto en términos de las cargas como de las comunalidades para las seis variables.
Con estos resultados, que pueden verse en el anexo 2 , ganamos en seguridad en cuanto a la estabilidad de los resultados.

6.2 CASO 2. Aplicación del Análisis Discriminante.

Definición del Problema y los Objetivos.
El servicio de estudio de una compañía de seguros pretende aplicar la técnica de análisis discriminante en una investigación sobre la siniestralidad en la rama del automóvil.
Entre los siniestros ocurridos en un año se han seleccionado aleatoriamente 40 pólizas, clasificadas en dos grupos, separando a los asegurados que han tenido un siniestro grave de los restantes. Para cada una de las pólizas se ha considerado la información sobre la edad del conductor, la antigüedad del vehículo y la potencia del mismo.
La base de datos de referencia [Pérez, 1997] se muestra en el anexo 3.

En un análisis previo de los datos, la inspección visual de los gráficos de caja para cada variable (figura 6.2) muestra que pudiera haber diferencias entre los grupos en términos del tipo de siniestralidad.

Supuestos básicos.
Los principales supuestos del Análisis discriminante son los relativos a la función discriminante (normalidad, linealidad y homocedasticidad) y a la estimación de la función

discriminante (matrices de varianza y covarianza iguales). Para los objetivos de este ejemplo consideraremos cumplidos estos supuestos (en el capitulo 3 puede verse más información sobre este tema).

Se trata pues de analizar cuales son las variables que contribuyen en mayor grado a discriminar a los sujetos en los dos grupos establecidos a priori. Para ello, estas variables que mejor discriminan se reducen a variables canónicas, que no son otra cosa sino una combinación lineal de las variables independientes originales. Esta combinación lineal es lo que se conoce como función discriminante, donde la variable dependiente es la pertenencia a un grupo u otro (siniestralidad grave o no) Como ya conocemos habrá tantas funciones discriminantes como grupos – 1 (k – 1) y para que sea óptima debe proporcionar una regla de clasificación que minimice la probabilidad de cometer errores.

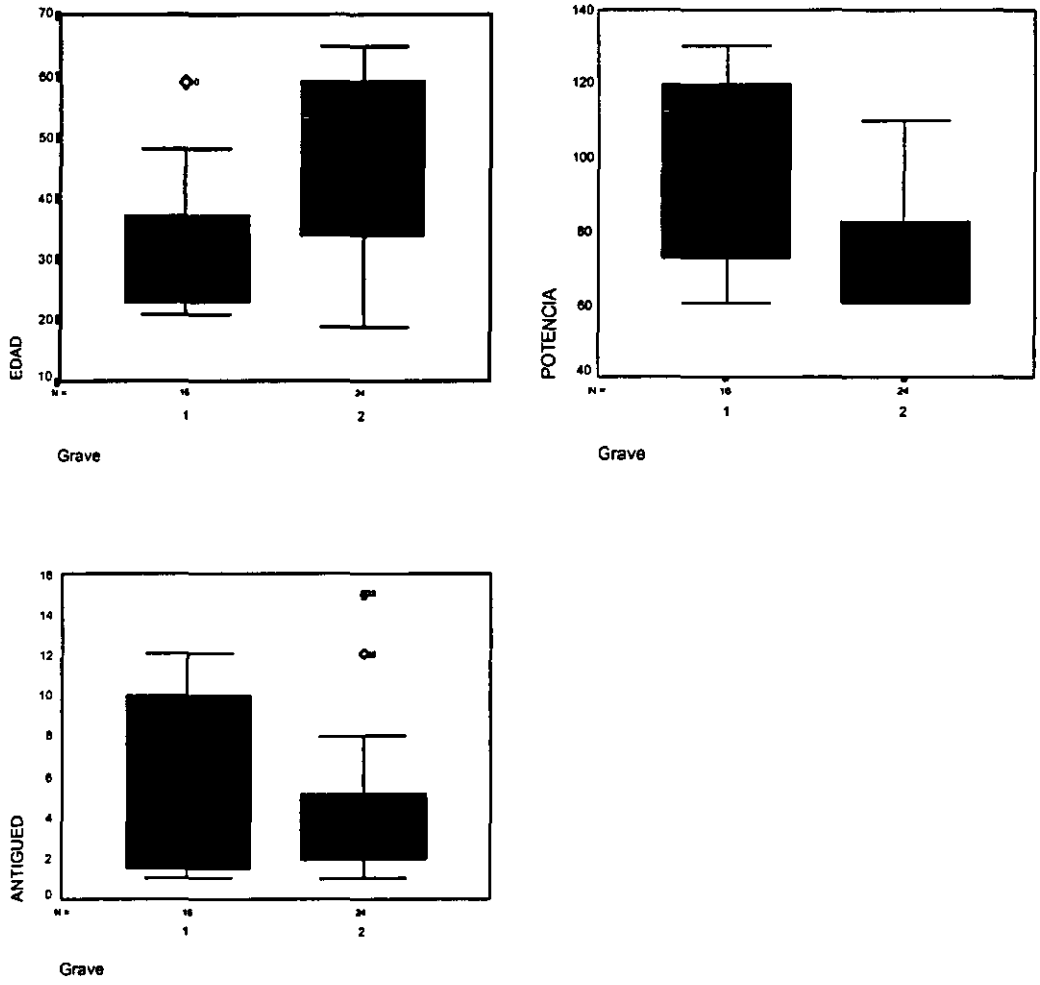


Figura 6.2 Gráficos de caja

Al igual que en el caso anterior el procesamiento se realizó con el paquete estadístico SPSS versión 8.0

Para iniciar cargamos los datos y entramos en *Statistics/Classify/Discriminant*. Nos aparece el cuadro de diálogo principal de la figura 6.3 *Discriminant Analysis*

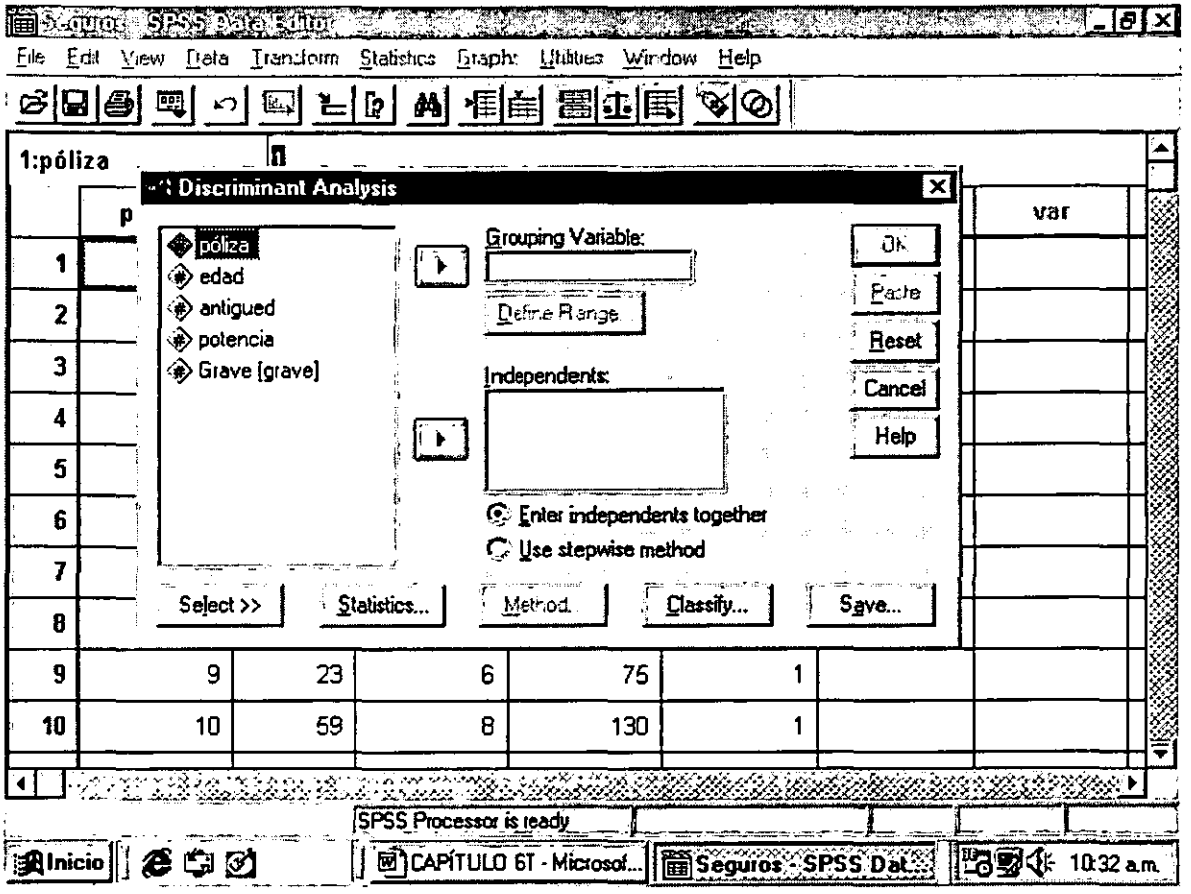


Figura 6.3 Cuadro de diálogo principal y subcuadros del análisis discriminante. SPSS 8.0

Comentaremos a continuación las diferentes opciones seleccionadas en este cuadro de diálogos, para después ejecutar el procedimiento.

Grouping Variable: Entramos en este campo la variable dependiente (grave) que es una variable con dos categorías (1= grave, 2 = no grave). En *Define Range* debemos especificar estos dos valores en mínimo y máximo.

Independients: Entramos las variables que actúan en el modelo como independientes o predictoras (edad, antigüedad, potencia).

Enter Independents Together or Use Stepwise Method: El modelo trabaja con el método simultaneo o con el método por etapas.
Dejamos la opción por defecto que es el simultaneo. (*Enter Independents Together*)

Select: Nos permite reducir el análisis a un subgrupo de la muestra total. Dejamos esta opción en blanco que es la opción por defecto.

Statistics: En este subcuadro seleccionamos todas las opciones.

Method: Sólo para cuando se selecciona use stepwise method. Que no es el caso.

Classification: En este subcuadro dejamos las opciones por defecto y seleccionamos el resto a excepción de *Limit case to first, Leave-one-out cassification* y *Replace missing values with mean.*

Save: Seleccionamos todas las opciones.

Después de esta selección pulsamos OK en el cuadro principal para ejecutar el procedimiento y obtenemos los siguientes resultados.

Discriminant

Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		40	100.0
Excluded	Missing or out-of-range group codes	0	.0
	At least one missing discriminating variable	0	.0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	.0
	Total	0	.0
Total		40	100.0

Group Statistics

Grave		Mean	Std. Deviation	Valid (listwise Unweight	Weighted
1	EDAD	31.19	10.86	1	16.000
	ANTIGUED	6.44	4.08	1	16.000
	POTENCIA	89.38	26.76	1	16.000
2	EDAD	43.79	14.50	2	24.000
	ANTIGUED	5.21	4.23	2	24.000
	POTENCIA	75.21	18.09	2	24.000
Total	EDAD	38.75	14.44	4	40.000
	ANTIGUED	5.70	4.16	4	40.000
	POTENCIA	80.88	22.76	4	40.000

La tablas Analysis Case Processing Summary nos deja ver que el análisis se realizó para las 40 pólizas seleccionadas.
Tenemos posteriormente la media y la desviación estándar de las variables, por separado para el grupo de pólizas con siniestralidad grave y no grave, y para la muestra total; donde se aprecia bastante diferencia en las medias para las tres variables en cada grupo.

En la siguiente tabla de salida Tests of Equality of Group Means tenemos la λ de Wilks y el anova univariante, utilizados para valorar la significación entre las medias de la variable independiente para los grupos.
Encontramos la diferencia más significativa en las variables edad con una F=8.775 y un grado de significación de 0.005, en segundo lugar la potencia con F= 4.008 y significación de 0.052, la antigüedad resulta no significativa.

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df	Sig.
EDAD	.812	8.775	1	3	.005
ANTIGUED	.979	.833	1	3	.367
POTENCIA	.905	4.008	1	3	.052

Tenemos a continuación la matriz de covarianza intragrupos, la matriz de correlación de Pearson y más adelante la matriz de covarianza para siniestros graves, no graves y total.

Pooled Within-Groups Matrices

		EDAD	ANTIGUED	POTENCIA
Covariance	EDAD	173.800	27.703	99.287
	ANTIGUED	27.703	17.418	13.904
	POTENCIA	99.287	13.904	480.729
Correlation	EDAD	1.000	.504	.343
	ANTIGUED	.504	1.000	.152
	POTENCIA	.343	.152	1.000

a The covariance matrix has 38 degrees of freedom.

Covariance Matrices

Grave		EDAD	ANTIGUED	POTENCIA
1	EDAD	117.896	14.046	80.458
	ANTIGUED	14.046	16.663	52.292
	POTENCIA	80.458	52.292	716.250
2	EDAD	210.259	36.611	111.567
	ANTIGUED	36.611	17.911	-11.132
	POTENCIA	111.567	-11.132	327.129
Total	EDAD	208.449	23.179	52.788
	ANTIGUED	23.179	17.344	17.833
	POTENCIA	52.788	17.833	517.804

a The total covariance matrix has 39 degrees of freedom.

Estimación de la función discriminante.

El siguiente resultado es el del test de Box cuyos valores aparecen a continuación.

Analysis 1

Box's Test of Equality of Covariance Matrices

Log Determinants

Grave	Rank	Log Determinant
1	3	13.768
2	3	13.025
Pooled within-groups	3	13.772

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Test Results

Box's M		17.247
F	Approx	2.611
	df	6
	df	6917.670
	Sig	.016

Tests null hypothesis of equal population covariance matrices.

Este test contrasta hasta que punto las matrices de varianza covarianza para cada grupo puede o no proceder de la misma población, es decir, si difieren o no significativamente.
En nuestro caso se aprecia una diferencia significativa con una f = 2.611 y un grado de significación de 0.016.

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalu	% of Variance	Cumulative %	Canonical Correlation
1	.71	100.0	100.0	.647

a First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.582	19.778	3	.000

Standardized Canonical Discriminant Function Coefficients

	Function
	1
EDAD	-1.116
ANTIGUED	.635
POTENCIA	.670

Structure Matrix

	Function
	1
EDAD	-.567
POTENCIA	.383
ANTIGUED	.175

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions Variables ordered by absolute size of correlation within function.

Canonical Discriminant Function Coefficients

	Function
	1
EDAD	-.085
ANTIGUED	.152
POTENCIA	.031
(Constant)	-.057

Unstandardized coefficients

De los resultados de salida Summary of Canonical Discriminant Functions obtenemos que la función discriminante es:

$$d = -0.057 - 0.085 \, x \, Edad + 0.152 \, x \, Antigüedad + 0.031 \, x \, Potencia$$

Multiplicando cada uno de estos coeficientes por los valores de cada póliza en las variables independientes obtenemos la puntuación discriminante para cada una de ellas. En el caso de la póliza 1 la puntuación discriminante sería:

$d_1 = - \ 0.057 - 0.085 \ (25) + 0.152 \ (6) + 0.031 \ (71)$

$d_1 = 5.181$

Este valor aparece al final de la base de datos en la segunda de las columnas generadas por el programa y correspondiente a las puntuaciones discriminantes.

Para estudiar que variable explica mejor la pertenencia a un grupo según la gravedad del siniestro se trabaja con los coeficientes estandarizados.

Variable	Coeficiente estandarizado
Edad	-1.1164
Antigüedad	0.6349
Potencia	0.6699

Se pone de manifiesto que la variable que más contribuye a explicar la pertenencia a un grupo es la edad del conductor, seguida en niveles muy similares por la potencia y la antigüedad del vehículo.

A partir de la puntuación discriminante es posible obtener una regla de clasificación de los casos en uno de los dos grupos.

La siguiente tabla recoge los centroides de los grupos, como puede apreciarse existe suficiente separación entre los mismos, indicativo de diferencia entre los grupos.

Functions at Group Centroids

	Functio
Grave	
1	1.01
2	-.67

Unstandardized canonical discriminant functions evaluated at group means

De manera general si el valor discriminante de una póliza d_i está más cerca del centroide del grupo I, se asignará a este grupo y en caso contrario al grupo II. Si tomamos

$$c = \frac{centroide\ \ I + centroide\ \ II}{2}$$

La regla de clasificación puede establecerse en los siguientes términos:

Si para la póliza i, $d_i < c \rightarrow$ Grupo II (no grave)
en caso contrario $\rightarrow$ Grupo I (grave)

Por otro lado el programa nos proporciona una regla de clasificación basada la teoría de Bayes y las probabilidades de que una póliza pertenezca a un grupo u otro.

Classification Statistics

Prior Probabilities for Groups

	Prior	Cases Used in Analysis	
Grave		Unweighted	Weighted
1	.500	16	16.000
2	.500	24	24.000
Total	1.000	40	40.000

Classification Function Coefficients

	Grave	
	1	2
EDAD	6.071E-02	.204
ANTIGUED	.138	-.119
POTENCIA	.169	.118
(Constant)	-9.653	-9.273

Fisher's linear discriminant functions

Los resultados de la clasificación aparece en la tabla Case Wise Statistics. En ella las columnas mas importantes son las siguientes.

Case number: Número de la póliza, en nuestro caso son en total 40
Actual Group: Grupo al que pertenecen (grave, no grave)
Predicted group: Grupo al que son asignadas de acuerdo a la función discriminadora.
Discriminant Score: Puntuación discriminante (Z)

Casewise Statistics

	Case Number	Actual Group	Highest Group					Second Highest Group			Discriminant Scores
			Predicted Group	P(D>d |G=g)		P(G=g | D=d)	Squared Mahalanobis Distance to Centroid	Group	P(G=g | D=d)	Squared Mahalanobis Distance to Centroid	Function 1
				p	df						
Original	1	1	1	.721	1	.883	.127	2	.117	4.179	1.369
	2	1	1	.719	1	.884	.129	2	.116	4.189	1.372
	3	1	1	.425	1	.519	.636	2	.481	.791	.215
	4	1	1	.944	1	.824	.005	2	.176	3.089	1.083
	5	1	1	.647	1	.657	.209	2	.343	1.512	.555
	6	1	1	.511	1	.578	.433	2	.422	1.060	.355
	7	1	1	.013	1	.996	6.134	2	.004	17.338	3.489
	8	1	2	.997	1	.807	.000	1	.193	2.860	-.679
	9	1	1	.851	1	.851	.035	2	.149	3.515	1.200
	10	1	2	.417	1	.514	.658	1	.486	.768	.136
	11	1	2	.568	1	.613	.326	1	.387	1.246	-.104
	12	1	1	.062	1	.990	3.480	2	.010	12.622	2.878
	13	1	2	.544	1	.599	.367	1	.401	1.169	-.069
	14	1	1	.760	1	.713	.093	2	.287	1.911	.707
	15	1	1	.418	1	.514	.657	2	.486	.769	.202
	16	1	1	.013	1	.996	6.134	2	.004	17.338	3.489
	17	2	1	.554	1	.604	.351	2	.396	1.199	.420
	18	2	2	.676	1	.672	.174	1	.328	1.613	-.258
	19	2	2	.961	1	.793	.002	1	.207	2.683	-.626
	20	2	2	.480	1	.558	.499	1	.442	.962	.032
	21	2	2	.969	1	.816	.002	1	.184	2.980	-.714
	22	2	1	.934	1	.783	.007	2	.217	2.574	.930
	23	2	2	.792	1	.866	.069	1	.134	3.804	-.938
	24	2	2	.452	1	.937	.566	1	.063	5.953	-1.428
	25	2	2	.444	1	.938	.585	1	.062	6.012	-1.440
	26	2	2	.494	1	.929	.467	1	.071	5.620	-1.358
	27	2	2	.539	1	.921	.378	1	.079	5.298	-1.289

	28	2	2	.619	1	.906	.247	1	.094	4.770	-1.172
	29	2	2	.544	1	.599	.367	1	.401	1.169	-.069
	30	2	1	.903	1	.836	.015	2	.164	3.274	1.134
	31	2	2	.694	1	.890	.155	1	.110	4.330	-1.068
	32	2	2	.444	1	.938	.585	1	.062	6.012	-1.440
	33	2	2	.822	1	.859	.051	1	.141	3.657	-.900
	34	2	2	.650	1	.659	.206	1	.341	1.520	-.221
	35	2	2	.539	1	.921	.378	1	.079	5.298	-1.289
	36	2	2	.619	1	.906	.247	1	.094	4.770	-1.172
	37	2	2	.400	1	.501	.707	1	.499	.716	.166
	38	2	2	.966	1	.794	.002	1	.206	2.705	-.632
	39	2	2	.452	1	.937	.566	1	.063	5.953	-1.428
	40	2	2	.444	1	.938	.585	1	.062	6.012	-1.440

** Misclassified case

Finalmente en la tabla Classification Result conocida también como matriz de confusión tenemos un resumen de la clasificación y se especifica el número de individuos correcta e incorrectamente clasificados del total de la muestra utilizada en el análisis.

Validación
Podemos constatar que el análisis discriminó correctamente 12 + 21 = 33 pólizas (diagonal principal) que representa el 82.5% . Puesto que una clasificación aleatoria se sitúa en un 50% de probabilidad podemos concluir que la función discriminadora es efectiva

Classification Results

			Predicted Group Membership		Total
		Grave	1	2	
Original	Count	1	12	4	16
		2	3	21	24
	%	1	75.0	25.0	100.0
		2	12.5	87.5	100.0

a 82.5% of original grouped cases correctly classified.

6.3 Caso 3. Aplicación del análisis de Grupos (*Cluster*).

Definición del Problema y los Objetivos.
Se dispone de información acerca de 14 provincias cubanas (datos no reales) sobre un conjunto de variables económicas. Para cada provincia se conocen los siguientes datos (variables métricas).

1. Producto interno bruto per cápita
2. Tasa de natalidad (número de nacimientos por 1000 habitantes)
3. Empleos en la agricultura (% sobre la población activa)
4. Empleo en la industria (% sobre la población activa)
5. Empleo en el sector de servicios (% sobre la población activa)

Provincias.

1	Las Tunas
2	Ciego de Avila
3	Santiago de Cuba
4	Cienfuegos
5	Pinar del Río
6	Matanzas
7	La Habana
8	Holguín
9	Ciudad Habana
10	Camaguey
11	Santi Spíritus
12	Guantanámo
13	Villa Clara
14	Granma

Se desea realizar la agrupación de las provincias mediante análisis de grupos de forma tal que sea posible tomar decisiones económicas para grupos de ellas.
Los datos de referencia aparecen en el anexo 4

Supuestos.
Las exigencias de normalidad, homocedasticidad y linealidad que son importantes en otras técnicas tienen poco peso en esta técnica.

Empleando el programa SPSS 8.0 se carga el archivo de datos y se entra en *Statistics/Classiffy/Hierarchical* Cluster y obtenemos el cuadro de dialogo principal del análisis de grupos jerárquico (figura 6.4). Se selecciona como variable a agrupar a las provincias y el resto de las variables se pasan como variables criterios de clasificación.

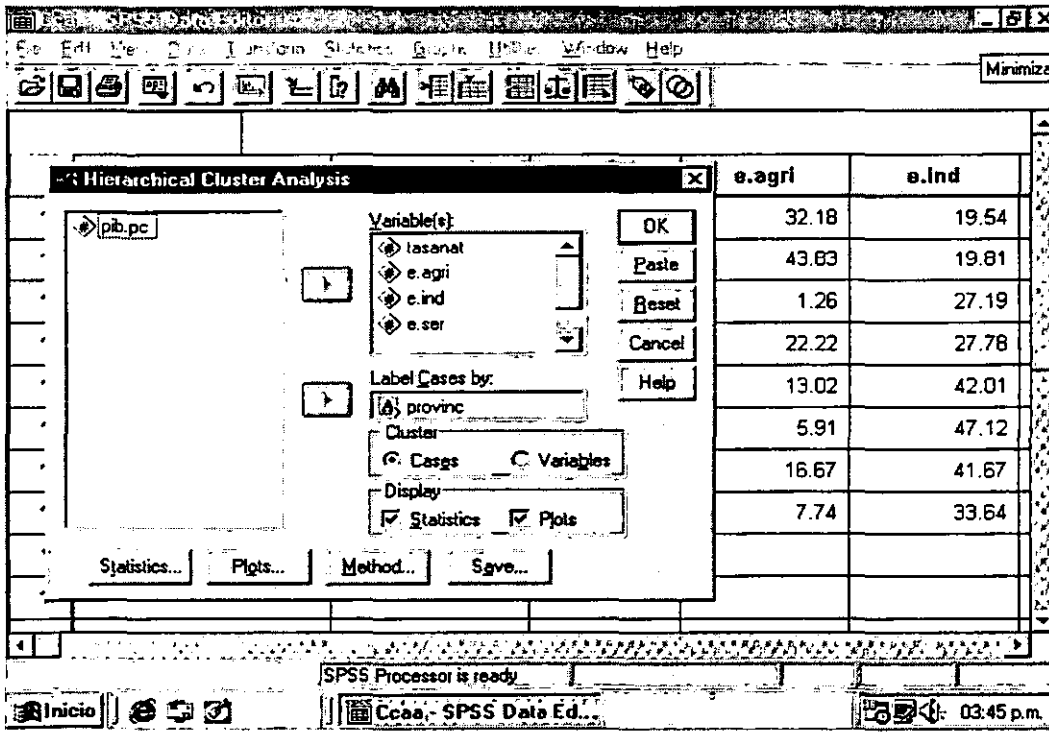


Figura 6.4 Cuadro de diálogo principal del análisis de grupos. Programa SPSS versión 8.0

Se selecciona en los diferentes subcuadros (*statistics, plots, method, save*) los gráficos y diseño de los diferentes métodos y valores de salida a obtener y que más adelante serán comentados.
Subcuadro Estadísticos (*Statistics)*: En este subcuadro se selecciona como salida la tabla de aglomeración y la matriz de distancias y también da la posibilidad de obtener la relación de los grupos que se van formando en cada paso (*cluster membership)*
Subcuadro Gráficos (*Plots*): Aquí podemos seleccionar la salida de gráficos de carámbanos y dendogramas.
Subcuadro Métodos (*Method*): Relaciona para ser seleccionado alguno de los métodos empleados para combinar los grupos, así como el tipo de medida de distancia a emplear y si se desea o no algún tipo de estandarización de las variables.
Subcuadro Salvar (*Save*): Esta opción permite grabar o no en la base de datos el cluster al que pertenece cada sujeto para un número determinado de cluster que sean seleccionados.

En el ejemplo que nos acupa definimos un análisis de grupos jerárquico aglomerativo ó ascendente, utilizando como medida de distancia la euclideana al cuadrado (datos métricos) y como método para la formación de los grupos el de enlace promedio (*between groups linkage*).
Con estas especificaciones obtenemos las siguientes salidas.
El primer resultado que obtenemos es la matriz de coeficientes de distancias euclideanas al cuadrado entre las diferentes provincias (*Proximity Matrix*).
Estos coeficientes nos indican que para una provincia determinada y en función de las variables criterio seleccionadas, cuanto mayor sea el coeficiente más diferencia existe con respecto a las otras provincias y viceversa.

A partir de la determinación de estos coeficientes (D^2) se van formando los grupos. En el gráfico de carámbanos vertical (*vertical icecle*) las columnas representan a cada provincia. Este gráfico se lee de abajo a arriba, por lo que la fila 13 y última representa el primer grupo y la fila 1 el último. Las provincias de Guantánamo y Las Tunas forman el primer grupo, el siguiente lo forman Camaguey y Holguín, el tercero lo constituyen Villa Clara y Guantánamo y así sucesivamente hasta el último grupo en el que están todas las provincias.

El número de filas se corresponde con el número de grupos a ese nivel, por lo que, por ejemplo, si trazamos una raya horizontal en la línea 3 cortamos dos barras blancas que indican los tres grupos a ese nivel.

Proximity Matrix

Case	Squared Euclidean Distance 1:Ciego de Avila	2:Las Tunas	3:Santiago	4:Cienfuegos	5:P.Rio	6:Matanzas	7:La Habana	8:Holguin	9:C.Habana	10:Camaguey	11:Santi Spiritus	12:Guantánamo	13:Villa Clara	14:Granma
1		356012457984.000	1119382798336.000	18625193984.000	80798392320.000	122321494016.000	861288333312.000	35080536064.000	894953259008.000	33950296064.000	53157796249.6000	355586703360.000	37661029171.2000	29423956787.2000
2	356012457984.000		212835581952.000	211778191360.000	97604526080.000	60970917888.000	109818281984.000	167583645696.000	122046750720.000	170083450880.000	17536677888.000	127784.578	289613120.000	2940812288.000
3	1119382798336.000	212835581952.000		849226235904.000	598701899776.000	501637840896.000	16887693312.000	758137421824.000	12541575168.000	763443937280.000	10818504294.4000	213165015040.000	19742297292.8000	26581275443.2000
4	18625193984.000	211778191360.000	849226235904.000		21837871104.000	45484412928.000	626602147840.000	2583062528.000	655363997696.000	2283111424.000	35119821619.2000	211449839616.000	22773098086.4000	16480709836.8000
5	80798392320.000	97604526080.000	598701899776.000	21837871104.000		4289507072.000	414485708800.000	9399804928.000	437938356224.000	9998902272.000	19788560793.6000	97381646336.000	10852758323.2000	66661031936.000
6	122321494016.000	60970917888.000	501637840896.000	45484412928.000	4289507072.000		334443970560.000	26389002240.000	355543646208.000	27386556416.000	14390568550.4000	60794793984.000	69664808960.000	37130842112.000
7	861288333312.000	109818281984.000	16887693312.000	626602147840.000	414485708800.000	334443970560.000		548722769920.000	322687168.000	553238659072.000	39585992704.000	110054948864.000	98828746752.000	14870098739.2000
8	35080536064.000	167583645696.000	758137421824.000	2583062528.000	9399804928.000	26389002240.000	548722769920.000		575658655744.000	9254000.000	29354285465.6000	167291568128.000	18180659609.6000	12612482662.4000
9	894953259008.000	122046750720.000	12541575168.000	655363997696.000	437938356224.000	355543646208.000	322687168.000	575658655744.000		580283858944.000	47056789504.0000	122296238080.000	11044580556.8000	16287775129.6000
10	33950296064.000	170083450880.000	763443937280.000	2283111424.000	9998902272.000	27386556416.000	553238659072.000	9254000.000	580283858944.000		29684829388.8000	169789194240.000	18440991539.2000	12829469081.6000
11	531577962496.000	17536677888.000	108185042944.000	351198216192.000	19788560793.6000	14390568550.4000	39585992704.000	2935428546.56000	47056789504.0000	2968482938.88000		17631332352.000	13319030784.000	34840231936.000
12	355586703360.000	127784.578	213165015040.000	211449839616.000	97381646336.000	60794793984.000	110054948864.000	167291568128.000	122296238080.000	169789194240.000	17631332352.000		301887936.000	2902231040.000
13	37661029171.2000	289613120.000	19742297292.8000	22773098086.4000	10852758323.2000	69664808960.000	98828746752.000	18180659609.6000	11044580556.8000	18440991539.2000	13319030784.000	301887936.000		5076174336.000
14	29423956787.2000	2940812288.000	26581275443.2000	16480709836.8000	66661031936.000	37130842112.000	14870098739.2000	12612482662.4000	16287775129.6000	12829469081.6000	34840231936.000	2902231040.000	5076174336.000	

This is a dissimilarity matrix

Vertical Icicle

Number of clusters	Case															
	9:C.Habana		7:La Habana		3:Santiago		11:Santi Spiritus		14:Granma		13:Villa Clara		12:Guantánamo		2:Las Tunas	
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
3	X	X	X	X	X		X	X	X	X	X	X	X	X	X	
4	X	X	X	X	X		X	X	X	X	X	X	X	X	X	
5	X	X	X	X	X		X	X	X	X	X	X	X	X	X	
6	X	X	X	X	X		X		X	X	X	X	X	X	X	
7	X	X	X		X		X		X	X	X	X	X	X	X	
8	X	X	X		X		X		X	X	X	X	X	X	X	
9	X	X	X		X		X		X		X	X	X	X	X	
10	X	X	X		X		X		X		X	X	X	X	X	
11	X		X		X		X		X		X	X	X	X	X	
12	X		X		X		X		X		X		X	X	X	
13	X		X		X		X		X		X		X	X	X	

Vertical Icicle

Number of clusters	Case										
	6:Matanzas		5:P.Rio		10:Camaguey		8:Holguin		4:Cienfuegos		1:Ciego de Avila
1	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X	X
3	X	X	X	X	X	X	X	X	X	X	X
4	X	X	X	X	X	X	X	X	X		X
5	X	X	X		X	X	X	X	X		X
6	X	X	X		X	X	X	X	X		X
7	X	X	X		X	X	X	X	X		X
8	X		X		X	X	X	X	X		X
9	X		X		X	X	X	X	X		X
10	X		X		X	X	X		X		X
11	X		X		X	X	X		X		X
12	X		X		X	X	X		X		X
13	X		X		X		X		X		X

La tabla de aglomeración (*Agglomerations Schedule*) que a continuación se muestra constituye otra de las salidas del programa que identifica los grupos que se van formando en cada paso. En esta tabla se relacionan además los coeficientes D^2 correspondientes. Las dos columnas siguientes, nos indican en que paso se forma por primera vez un multigrupo, o sea, un grupo en el que a uno ya existente se le adiciona un tercer elemento (ejemplo en el paso 3 al grupo formado desde el paso 1 entre las provincias de Las Tunas y Guantánamo se le une la provincia de Villa Clara). Por último en la columna final, se refleja el paso en el que al grupo formado se le unirá un nuevo componente.

Agglomeration Schedule

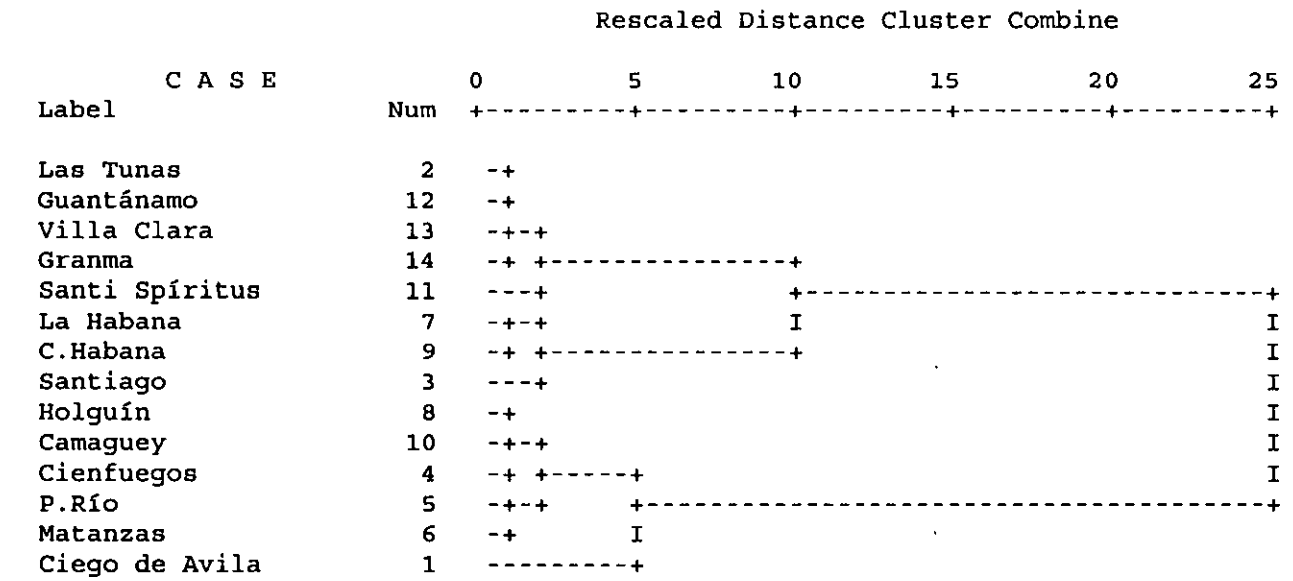
	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
Stage	Cluster	Cluster 2		Cluster 1	Cluster 2	
1		12	127784.578	0	0	3
2		10	9254000.000	0	0	5
3		13	295750528.000	1	0	6
4		9	322687168.000	0	0	8
5		8	2433086976.000	0	2	10
6		14	3639739136.000	3	0	9
7		6	4289507072.000	0	0	10
8		7	14714634240.000	0	4	12
9		11	20831817728.000	6	0	12
10		5	23416092672.000	5	7	11
11		4	58155180032.000	0	10	13
12		3	137942237184.000	9	8	13
13		2	361173909504.000	11	12	0

Finalmente se obtiene el dendograma resultante del análisis. Este gráfico se lee de izquierda a derecha y las líneas verticales representan la unión formando nuevos grupos. En la parte superior aparece una escala y la posición de la línea vertical respecto a esta escala indica la distancia a la cual los grupos se unen.

* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * * *

Dendrogram using Average Linkage (Between Groups)

Rescaled Distance Cluster Combine



Finalmente tenemos la siguiente solución en tres grupos.

Grupo 1	Grupo 2	Grupo 3
Las Tunas	Matanzas	Ciudad Habana
Guantánamo	Pinar del Río	La Habana
Villa Clara	Camaguey	Santiago de Cuaba
Granma	Holguín	
Santi Spíritus	Cienfuegos	
	Ciego de Ávila	

Con vistas a validar estos resultados se llevó a cabo la aplicación de un método de análisis de grupos alternativo: el Cluster no jerárquico (*K- Medias*) para 3 grupos, con centros desconocidos, basado en la asignación de cada elemento a los grupos con bases en que su distancia con respecto al centro del mismo sea mínima.

Todas las salidas de este procedimiento se muestran en el anexo 6 y se destaca en la tabla *Cluster Membership* la relación de las provincias, el grupo a que fueron asignadas y la distancia euclideana respecto al centro de los mismos.
Se puede apreciar que la clasificación coincide totalmente con la obtenida con el procedimiento inicial.

Cluster Membership

Case Number	Provincia	Cluster	Distance
1	Las Tunas	1	18971.403
2	Ciego de Avila	2	36262.001
3	Santiago	3	80647.667
4	Cienfuegos	2	99183.000
5	P.Río	2	48593.000
6	Matanzas	2	114088.000
7	La Habana	3	49305.336
8	Holguín	2	48359.004
9	C.Habana	3	31342.335
10	Camaguey	2	51401.001
11	Santi Spíritus	1	113454.600
12	Guantánamo	1	19328.403
13	Villa Clara	1	1953.415
14	Granma	1	73201.401

Se obtuvieron además resultados del análisis ANOVA que nos permite valorar la variabilidad entre grupos (*cluster mean square*) y dentro de cada grupo (*error mean square*). De estos resultados se destaca que la variable tasa de natalidad no representa una gran diferencia entre los grupos, mientras que las restantes variables ofrecen una diferenciación significativa entre los grupos.

Finalmente los resultados alcanzados por los dos métodos nos confirman la consistencia de los resultados.

CONCLUCIONES Y RECOMENDACIONES.

Conclusiones.

1. El análisis multivariante es una herramienta poderosa que proporciona una gran variedad de alternativas en el análisis de datos.

2. Las técnicas de análisis multivariante pueden clasificarse según sea de interés distinguir entre variables explicativas y explicadas, según sea la naturaleza y número de las variables a estudiar y según el objetivo del estudio.

3. El análisis e interpretación de cualquier problema multivariante puede verse ayudado por una metodología, que ofrece una serie de pautas a seguir en su aplicación.

4. La metodología parte de la definición conceptual del problema y los objetivos, lo que debe constituir el primer paso en cualquier análisis multivariante, continua con el análisis previo de los datos, y la evaluación de los supuestos básicos, la identificando las relaciones fundamentales a estudiar o modelización y solo cuando el modelo ya interpretado ha superado la etapa de validación se está en condiciones de llevar a cabo su implementación práctica.

5. La aplicación indiscriminada de técnicas estadísticas de análisis multivariante inadecuadas a cualquier base de datos y la explicación posterior de los resultados obtenidos construyendo supuestas teorías, es uno de los errores más frecuentes en el uso de estas técnicas en la actualidad.

6. Debido a su complejidad matemática, es prácticamente imposible la aplicación de estas técnicas sin la utilización de la computación y los paquetes de programas desarrollados con este propósito, por lo que un correcto entrenamiento en su utilización constituye una garantía para la obtención de resultados adecuados.

7. Existen una gran variedad de paquetes de programas creados para el análisis estadístico y en particular para el análisis multivariante. Cada investigador deberá escoger el apropiado para su caso y aquel que le resulte más cómodo para su uso.

Recomendaciones.

1. Dado que la gran mayoría de los fenómenos a que nos enfrentamos en nuestras vidas están influidos por la acción de muchas variables de forma simultanea, en su estudio se recomienda la aplicación de técnicas de análisis multivariante que tengan en cuenta este comportamiento. De otra manera sólo estaremos considerando parte del fenómeno

2. El éxito en el análisis multivariante implica mucho más que la selección del método correcto, por lo que de una forma u otra deben ejecutarse los pasos incluidos en la metodología propuesta para obtener resultados aplicables.

3. Para una mejor aplicación de los diferentes Software se recomienda consultar los manuales editados con este propósito.

Bibliografía.

Alvarez, R. Manual de utilización del SPAD-Windows. Traducción al español. *ISPJAE, Cuba - Université Libre de Bruxelles, Belgique. Programa PRESTA.*

Anderson, T.W. (1984), An introduction to multivariate statistical analysis. *John Wiley & Sons, INC. Second Edition.*

Calvo, F. (1993), Técnicas de estadística multivariante. *Bilbao.: Universidad de Duesto.*

Carroll, J y Green, P. (1997), Mathematical tools for applied multivariate analysis. *Editorial Academic* Press.

Cuadras, C. M. (1991), Métodos de análisis multivariante. *Barcelona. Eunibar, Segunda Edición.*

Dilon,W. R. (1984), Multivariate análisis. Methods and Applications. *John Wiley & Sons. INC.*

Etxeberria, J; Joaristi, L y Lizasoain, L. (1990), Programación y análisis estadístico básico con SPSS-PC(+). *Madrid. Paraninfo.*

Everitt, B. S (1993), Cluster Analysis. *London: Edward Arnol*

Everitt, B.S y Dunn, G. (1991), Applied multivariate data analysis. *Edward Arnold a division of Hodder & Stoughton.*

Flury, B. (1997), A first Course in Multivariate Statistics, *Springer Text in Statistics.*

Flury, B y Riedwul, H. (1988), Multivariate Statistics. A practical approach. *Chapman and Hall.*

Freixa, I. B. (1992), Análisis exploratorio de datos. Nuevas técnicas estadísticas. *Barcelona. Ed. Promociones y Publicaciones Universitarias.*

Gnanadesikan, R. (1997), Methods for statistical data analysis of multivariate observations. *A. Wiley-Interscience publication, John Wiley & Sons. INC.*

Hair, J. F.; Anderson, R. E.; Tatham, R. L. y Black, W. C. (1999), Análisis Multivariante. *Prentice Hall, Quinta Edición.*

Hand, D. J. y Taylos, C. C. (1987), Multivariate analysis of variance and repeated measures.

Hardyck, C. D and L.F Petrinovich (1976), Introduction to statistics for Behavioral Sciences, *2d edition. Philadelphia: Saunders.*

Hernández, R; Fernández, C y Baptista, P. (1998), Metodología de la investigación. *McGraw Hill. Segunda Edición.*

Johnson, R. A. y Wichern, D. W. (1982), Applied multivariate statistical analysis. *Upper. Saddle River, NJ.: Prentice Hall.*

Journal of Marketing Research. Vol. 20, 1983 pp. 134 -148

Krzanowski, W. J. (1990), Principles of multivariate analysis a user perspective. C*laredon Press. Oxford.*

Krzanowski, W. J. y Marriot, F. (1994), Multivariate analysis. Distribution ordination and inference. Part I. *Kendall's Library of Statistics.*

Krzanowski, W. .J y Marriot, F. (1995), Multivariate analysis. Classification Structure and repeated Measurements. Part II. *Kendall's Library of Statistics.*

Krzanowski, W. J. (1995), Recent advances in descriptive multivariate analysis. *Oxford Science publications. Claredon Press, Oxford.*

Morrison, D. F. (1990), Multivariate statistical methods. *McGraw Hill. Third Edition, Series in Probability and Statistics.*

Norusis, M.J. (1995) The SPSS 6.0 guide to data analysis. USA: Prentice Hall.

Ott, Lyman. (1984), An introduction to statistical methods and data analysis. *Boston. Duxbury.*

Pérez, R. y López, A. J. (1997), Análisis de datos económicos II. Métodos Inferenciales. *Editorial Pirámides, S.A*

Siegel, A. F. y Morgan, C. J. (1996), Statistics and data analysis an introduction. *John Wiley & Sons. INC. Second Edition.*

Sprent, P. (1998), Data driven statistical methods. *Chapman & Hall.*

SPSS Inc. (1993), SPSS base system syntax reference guide, Release 6.0. *USA: SPSS, Inc.*

Stewart, D and Love, W. (1968), A General Canonical Correlation Index, Psychological Bolletin 70: pp. 160 - 163

Taylor, J. K. (1990), Statistical techniques for data analysis. Chelsea, *Michigan. Lewis.*

Uriel, E.(1995), Análisis de datos. Series temporales y análisis multivariante. *A.C.*

Van de Geer. (1971), Introduction to multivariate analysis for the social sciences. *University of Leiden, The Netherlands. Ed. WH Freeman and Company. San Francisco.*

Visauta Vinacua, B. (1998), Análisis Estadístico con SPSS para Windows. Estadística Multivariante Vol. II, *McGraw Hill.*

Watson, Billengsley,Craft and Huntsberger. (1993), Statistics for management and economics. *Ed. Ally and Bacon*

Anexos

Anexo 1. Base de datos atributos de la empresa.

	X1	X2	X3	X4	X5	X6	X7	Fact 1	Fact 2
1	4.1	0.6	6.9	4.7	2.4	2.3	5.2	-0.65395	-0.63487
2	1.8	3	6.3	6.6	2.5	4	8.4	1.17748	1.296
3	3.4	5.2	5.7	6	4.3	2.7	8.2	1.49277	0.52149
4	2.7	1	7.1	5.9	1.8	2.3	7.8	0.26456	-0.16299
5	6	0.9	9.6	7.8	3.4	4.6	4.5	-2.11554	2.56668
6	1.9	3.3	7.9	4.8	2.6	1.9	9.7	1.2621	-0.74199
7	4.6	2.4	9.5	6.6	3.5	4.5	7.6	-0.70641	2.00546
8	1.3	4.2	6.2	5.1	2.8	2.2	6.9	1.47782	-0.55378
9	5.5	1.6	9.4	4.7	3.5	3	7.6	-0.97737	0.22803
10	4	3.5	6.5	6	3.7	3.2	8.7	0.78749	0.82883
11	2.4	1.6	8.8	4.8	2	2.8	5.8	-0.34814	-0.30482
12	3.9	2.2	9.1	4.6	3	2.5	8.3	-0.15426	-0.27982
13	2.8	1.4	8.1	3.8	2.1	1.4	6.6	-0.0543	-1.6318
14	3.7	1.5	8.6	5.7	2.7	3.7	6.7	-0.56806	0.84786
15	4.7	1.3	9.9	6.7	3	2.6	6.8	-1.12896	0.72327
16	3.4	2	9.7	4.7	2.7	1.7	4.8	-0.86435	-0.91198
17	3.2	4.1	5.7	5.1	3.6	2.9	6.2	0.8828	0.08807
18	4.9	1.8	7.7	4.3	3.4	1.5	5.9	-0.57411	-1.09466
19	5.3	1.4	9.7	6.1	3.3	3.9	6.8	-1.29428	1.38239
20	4.7	1.3	9.9	6.7	3	2.6	6.8	-1.12896	0.72327
21	3.3	0.9	8.6	4	2.1	1.8	6.3	-0.52442	-1.25177
22	3.4	0.4	8.3	2.5	1.2	1.7	5.2	-0.78328	-2.04771
23	3	4	9.1	7.1	3.5	3.4	8.4	0.43421	1.44817
24	2.4	1.5	6.7	4.8	1.9	2.5	7.2	0.4512	-0.54484
25	5.1	1.4	8.7	4.8	3.3	2.6	3.8	-1.47794	-0.18527
26	4.6	2.1	7.9	5.8	3.4	2.8	4.7	-0.82787	0.37827

27	2.4	1.5	6.6	4.8	1.9	2.5	7.2	0.47611	-0.54841
28	5.2	1.3	9.7	6.1	3.2	3.9	6.7	-1.31176	1.35959
29	3.5	2.8	9.9	3.5	3.1	1.7	5.4	-0.60597	-1.3416
30	4.1	3.7	5.9	5.5	3.9	3	8.4	0.92149	0.4756
31	3	3.2	6	5.3	3.1	3	8	0.99289	0.20736
32	2.8	3.8	8.9	6.9	3.3	3.2	8.2	0.46649	1.17639
33	5.2	2	9.3	5.9	3.7	2.4	4.6	-1.355	0.27242
34	3.4	3.7	6.4	5.7	3.5	3.4	8.4	0.95251	0.75885
35	2.4	1	7.7	3.4	1.7	1.1	6.2	0.00335	-2.10995
36	1.8	3.3	7.5	4.5	2.5	2.4	7.6	0.93965	-0.6324
37	3.6	4	5.8	5.8	3.7	2.5	9.3	1.36143	0.26587
38	4	0.9	9.1	5.4	2.4	2.6	7.3	-0.71643	0.01928
39	0	2.1	6.9	5.4	1.1	2.6	8.9	1.50901	-0.40216
40	2.4	2	6.4	4.5	2.1	2.2	8.8	0.99266	-0.79959
41	1.9	3.4	7.6	4.6	2.6	2.5	7.7	0.92351	-0.49651
42	5.9	0.9	9.6	7.8	3.4	4.6	4.5	-2.08885	2.55461
43	4.9	2.3	9.3	4.5	3.6	1.3	6.2	-0.78152	-1.03206
44	5	1.3	8.6	4.7	3.1	2.5	3.7	-1.4618	-0.32117
45	2	2.6	6.5	3.7	2.4	1.7	8.5	1.20862	-1.48497
46	5	2.5	9.4	4.6	3.7	1.4	6.3	-0.77347	-0.88824
47	3.1	1.9	10	4.5	2.6	3.2	3.8	-1.18318	-0.0803
48	3.4	3.9	5.6	5.6	3.6	2.3	9.1	1.4179	0.00201
49	5.8	0.2	8.8	4.5	3	2.4	6.7	-1.38132	-0.37417
50	5.4	2.1	8	3	3.8	1.4	5.2	-0.81935	-1.65803
51	3.7	0.7	8.2	6	2.1	2.5	5.2	-0.88395	0.07529
52	2.6	4.8	8.2	5	3.6	2.5	9	1.17841	-0.06674
53	4.5	4.1	6.3	5.9	4.3	3.4	8.8	0.85692	1.01917
54	2.8	2.4	6.7	4.9	2.5	2.6	9.2	0.953	-0.2596
55	3.8	0.8	8.7	2.9	1.6	2.1	5.6	-0.84785	-1.50414
56	2.9	2.6	7.7	7	2.8	3.6	7.7	0.3191	1.34216

57	4.9	4.4	7.4	6.9	4.6	4	9.6	0.64913	1.98643
58	5.4	2.5	9.6	5.5	4	3	7.7	-0.77717	0.65001
59	4.3	1.8	7.6	5.4	3.1	2.5	4.4	-0.77766	-0.07349
60	2.3	4.5	8	4.7	3.3	2.2	8.7	1.20193	-0.47084
61	3.1	1.9	9.9	4.5	2.6	3.1	3.8	-1.15135	-0.1493
62	5.1	1.9	9.2	5.8	3.6	2.3	4.5	-1.33886	0.13653
63	4.1	1.1	9.3	5.5	2.5	2.7	7.4	-0.73329	0.16668
64	3	3.8	5.5	4.9	3.4	2.6	6	0.89782	-0.25706
65	1.1	2	7.2	4.7	1.6	3.2	10	1.30732	-0.15186
66	3.7	1.4	9	4.5	2.6	2.3	6.8	-0.55353	-0.58796
67	4.2	2.5	9.2	6.2	3.3	3.9	7.3	-0.512	1.37708
68	1.6	4.5	6.4	5.3	3	2.5	7.1	1.43611	-0.19658
69	5.3	1.7	8.5	3.7	3.5	1.9	4.8	-1.14101	-1.05941
70	2.3	3.7	8.3	5.2	3	2.3	9.1	0.99774	-0.22646
71	3.6	5.4	5.9	6.2	4.5	2.9	8.4	1.46049	0.79328
72	5.6	2.2	8.2	3.1	4	1.6	5.3	-0.89401	-1.44107
73	3.6	2.2	9.9	4.8	2.9	1.9	4.9	-0.91482	-0.6871
74	5.2	1.3	9.1	4.5	3.3	2.7	7.3	-0.93074	-0.13556
75	3	2	6.6	6.6	2.4	2.7	8.2	0.59061	0.51619
76	4.2	2.4	9.4	4.9	3.2	2.7	8.5	-0.23994	0.0517
77	3.8	0.8	8.3	6.1	2.2	2.6	5.3	-0.9001	0.21118
78	3.3	2.6	9.7	3.3	2.9	1.5	5.2	-0.57369	-1.61338
79	1	1.9	7.1	4.5	1.5	3.1	9.9	1.32525	-0.33185
80	4.5	1.6	8.7	4.6	3.1	2.1	6.8	-0.63189	-0.57301
81	5.5	1.8	8.7	3.8	3.6	2.1	4.9	-1.21567	-0.84245
82	3.4	4.6	5.5	8.2	4	4.4	6.3	0.86082	2.49567
83	1.6	2.8	6.1	6.4	2.3	3.8	8.2	1.20977	1.02421
84	2.3	3.7	7.6	5	3	2.5	7.4	0.82221	-0.25649
85	2.6	3	8.5	6	2.8	2.8	6.8	0.19011	0.37675
86	2.5	3.1	7	4.2	2.8	2.2	9	1.12791	-0.80551

87	2.4	2.9	8.4	5.9	2.7	2.7	6.7	0.23294	0.22879
88	2.1	3.5	7.4	4.8	2.8	2.3	7.2	0.85449	-0.52827
89	2.9	1.2	7.3	6.1	2	2.5	8	0.23228	0.1088
90	4.3	2.5	9.3	6.3	3.4	4	7.4	-0.55233	1.50504
91	3	2.8	7.8	7.1	3	3.8	7.9	0.3402	1.5542
92	4.8	1.7	7.6	4.2	3.3	1.4	5.8	-0.55797	-1.23055
93	3.1	4.2	5.1	7.8	3.6	4	5.9	0.8987	1.96417
94	1.9	2.7	5	4.9	2.2	2.5	8.2	1.49633	-0.49868
95	4	0.5	6.7	4.5	2.2	2.1	5	-0.62417	-0.88665
96	0.6	1.6	6.4	5	0.7	2.1	8.4	1.29435	-0.90475
97	6.1	0.5	9.2	4.8	3.3	2.8	7.1	-1.44159	0.1053
98	2	2.8	5.2	5	2.4	2.7	8.4	1.46834	-0.29099
99	3.1	2.2	6.7	6.8	2.6	2.9	8.4	0.60992	0.77232
100	2.5	1.8	9	5	2.2	3	6	-0.35374	-0.04511

Anexo 2. Validación Análisis factorial.

Submuestra 1. n = 50

Factor Analysis

Descriptive Statistics

	Mea	Std. Deviation	Analysis N
X1	3.64	1.350	50
X2	2.22	1.179	50
X3	8.02	1.417	50
X4	5.19	1.142	50
X6	2.61	.854	50
X7	6.83	1.575	50

Correlation Matrix

		X1	X2	X	X4	X6	X
Correlation	X	1.000	-.384	.53	.226	.231	-.50
	X	-.384	1.000	-.53	.182	.056	.54
	X	.531	-.531	1.00	.115	.134	-.52
	X	.226	.182	.11	1.000	.796	.16
	X	.231	.056	.13	.796	1.000	.05
	X	-.500	.545	-.52	.166	.050	1.00
Sig. (1-tailed)	X		.003	.00	.058	.054	.00
	X	.003		.00	.103	.350	.00
	X	.000	.000		.213	.177	.00
	X	.058	.103	.21		.000	.12
	X	.054	.350	.17	.000		.36
	X	.000	.000	.00	.125	.366	

a Determinant = 8.158E-02

Inverse of Correlation Matrix

	X	X2	X3	X	X6	X7
X1	1.71	.178	-.441	-.46	-.008	.602
X2	.17	1.720	.612	-.52	.221	-.448
X3	-.44	.612	1.841	-.38	.104	.477
X4	-.46	-.525	-.385	3.18	-2.317	-.560
X6	-.00	.221	.104	-2.31	2.811	.175
X7	.60	-.448	.477	-.56	.175	1.881

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.661
Bartlett's Test of Sphericity	Approx. Chi-Square	115.702
	df	15
	Sig.	.000

Anti-image Matrices

		X1	X2	X	X4	X6	X
Anti-image Covariance	X1	.585	6.061E-02	-.14	-8.468E-02	-1.768E-03	.18
	X2	6.061E-02	.581	.19	-9.593E-02	4.568E-02	-.13
	X3	-.140	.193	.54	-6.576E-02	2.004E-02	.13
	X4	-8.468E-02	-9.593E-02	-6.576E-0	.314	-.259	-9.362E-0
	X6	-1.768E-03	4.568E-02	2.004E-0	-.259	.356	3.306E-0
	X7	.187	-.139	.13	-9.362E-02	3.306E-02	.53
Anti-image Correlation	X1	.777	.104	-.24	-.197	-3.875E-03	.33
	X2	.104	.752	.34	-.224	.100	-.24
	X3	-.249	.344	.76	-.159	4.560E-02	.25
	X4	-.197	-.224	-.15	.497	-.775	-.22
	X6	-3.875E-03	.100	4.560E-0	-.775	.535	7.602E-0
	X7	.336	-.249	.25	-.229	7.602E-02	.74

a Measures of Sampling Adequacy(MSA)

Communalities

	Initia	Extraction
X1	1.00	.647
X2	1.00	.652
X3	1.00	.688
X4	1.00	.903
X6	1.00	.854
X7	1.00	.703

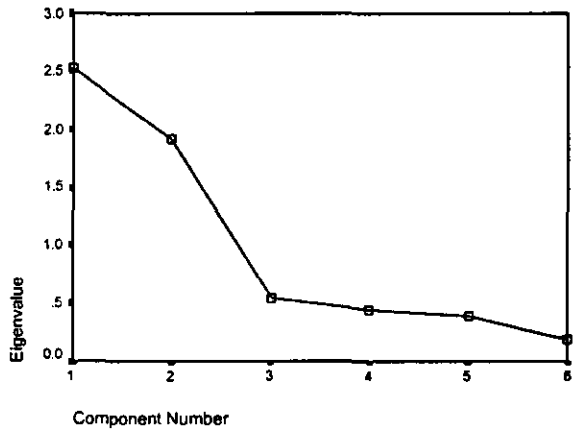
Extraction Method: Principal Component Analysis.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.527	42.108	42.108	2.527	42.108	42.108	2.513	41.879	41.879
2	1.920	32.001	74.109	1.920	32.001	74.109	1.934	32.231	74.109
3	.541	9.010	83.119						
4	.437	7.285	90.404						
5	.396	6.601	97.005						
6	.180	2.995	100.000						

Extraction Method: Principal Component Analysis.

Scree Plot



Component Matrix

	Component	
	1	2
X3	.829	2.889E-02
X7	-.791	.276
X1	.779	.197
X2	-.747	.306
X4	.102	.945
X6	.191	.904

Extraction Method: Principal Component Analysis.
a 2 components extracted.

Reproduced Correlations

		X1	X2	X	X4	X6	X
Reproduced Correlation	X1	.647	-.522	.65	.266	.327	-.56
	X2	-.522	.652	-.61	.213	.134	.67
	X3	.652	-.611	.68	.112	.184	-.64
	X4	.266	.213	.11	.903	.874	.18
	X6	.327	.134	.18	.874	.854	9.867E-0
	X7	-.562	.676	-.64	.180	9.867E-02	.70
Residual	X1		.138	-.12	-4.058E-02	-9.668E-02	6.216E-0
	X2	.138		8.008E-0	-3.077E-02	-7.826E-02	-.13
	X3	-.121	8.008E-02		2.912E-03	-5.036E-02	.12
	X4	-4.058E-02	-3.077E-02	2.912E-0		-7.795E-02	-1.411E-0
	X6	-9.668E-02	-7.826E-02	-5.036E-0	-7.795E-02		-4.905E-0
	X7	6.216E-02	-.131	.12	-1.411E-02	-4.905E-02	

Extraction Method: Principal Component Analysis.

a Residuals are computed between observed and reproduced correlations. There are 10 (66.0%) nonredundant residuals with absolute values > 0.05.

b Reproduced communalities

Rotated Component Matrix

	Component	
	1	2
X7	.824	.154
X3	-.815	.154
X2	.785	.190
X1	-.741	.313
X4	4.132E-02	.949
X6	-5.232E-02	.923

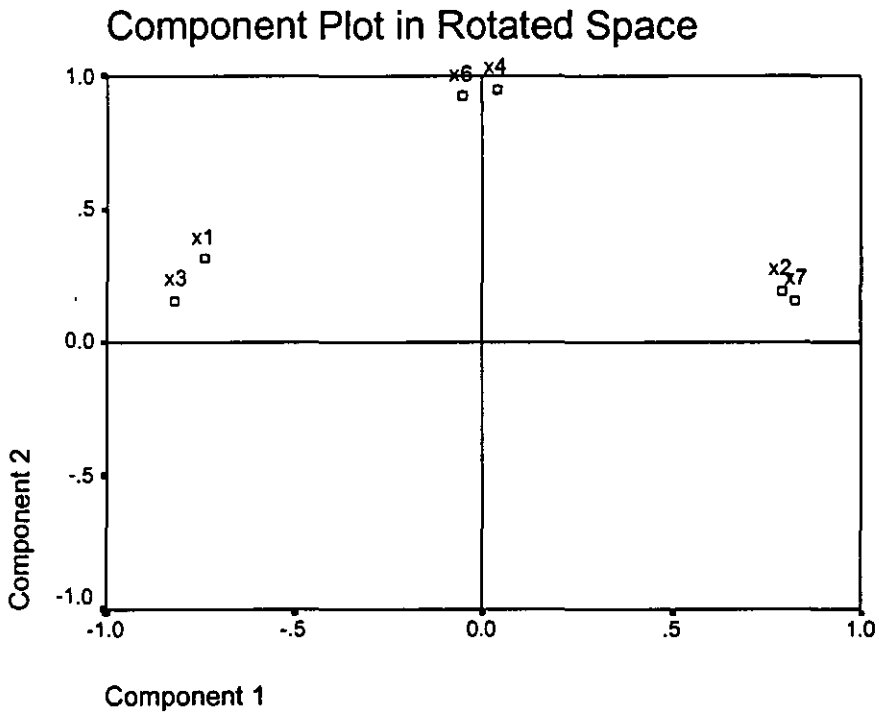
Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.

a Rotation converged in 3 iterations.

Component Transformation Matrix

Component	1	2
1	-.989	.151
2	.151	.989

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.



Component Score Coefficient Matrix

	Component	
	1	2
X1	-.289	.148
X2	.316	.113
X3	-.322	.064
X4	.034	.492
X6	-.004	.477
X7	.331	.095

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. Component Scores.

Component Score Covariance Matrix

Component	1	2
1	1.000	.000
2	.000	1.000

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. Component Scores.

Submuestra 2. n = 50

Factor Analysis

Descriptive Statistics

	Mea	Std. Deviation	Analysis N
X1	3.38	1.291	50
X2	2.50	1.207	50
X3	7.76	1.358	50
X4	5.29	1.130	50
X6	2.71	.683	50
X7	7.11	1.599	50

Correlation Matrix

		X1	X2	X	X4	X6	X
Correlation	X1	1.000	-.299	.47	-.125	-.108	-.45
	X2	-.299	1.000	-.43	.355	.336	.38
	X3	.476	-.432	1.00	-.353	-.242	-.35
	X4	-.125	.355	-.35	1.000	.787	.22
	X6	-.108	.336	-.24	.787	1.000	.32
	X7	-.456	.387	-.35	.228	.328	1.00
Sig. (1-tailed)	X1		.018	.00	.193	.228	.00
	X2	.018		.00	.006	.009	.00
	X3	.000	.001		.006	.045	.00
	X4	.193	.006	.00		.000	.05
	X6	.228	.009	.04	.000		.01
	X7	.000	.003	.00	.055	.010	

a Determinant = .125

Inverse of Correlation Matrix

	X	X2	X3	X	X6	X7
X1	1.49	.072	-.546	-.05	-.121	.510
X2	.07	1.406	.370	-.19	-.127	-.294
X3	-.54	.370	1.598	.54	-.275	.144
X4	-.05	-.193	.548	2.93	-2.215	.304
X6	-.12	-.127	-.275	-2.21	2.885	-.544
X7	.51	-.294	.144	.30	-.544	1.506

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.664
Bartlett's Test of Sphericity	Approx. Chi-Square	95.828
	df	15
	Sig.	.000

Anti-image Matrices

		X1	X2	X	X4	X6	X
Anti-image Covariance	X1	.669	3.428E-02	-.22	-1.189E-02	-2.812E-02	.22
	X2	3.428E-02	.711	.16	-4.690E-02	-3.128E-02	-.13
	X3	-.229	.165	.62	.117	-5.972E-02	5.982E-0
	X4	-1.189E-02	-4.690E-02	.11	.341	-.262	6.881E-0
	X6	-2.812E-02	-3.128E-02	-5.972E-0	-.262	.347	-.12
	X7	.226	-.139	5.982E-0	6.881E-02	-.125	.66
Anti-image Correlation	X1	.691	4.970E-02	-.35	-2.488E-02	-5.838E-02	.34
	X2	4.970E-02	.850	.24	-9.521E-02	-6.300E-02	-.20
	X3	-.353	.247	.72	.253	-.128	9.281E-0
	X4	-2.488E-02	-9.521E-02	.25	.582	-.762	.14
	X6	-5.838E-02	-6.300E-02	-.12	-.762	.575	-.26
	X7	.340	-.202	9.281E-0	.145	-.261	.71

a Measures of Sampling Adequacy(MSA)

Communalities

	Initia	Extraction
X1	1.00	.705
X2	1.00	.492
X3	1.00	.589
X4	1.00	.870
X6	1.00	.868
X7	1.00	.547

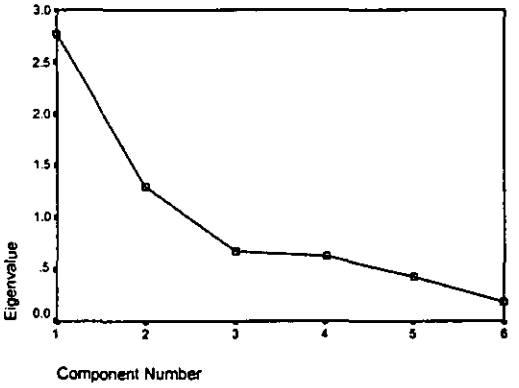
Extraction Method: Principal Component Analysis.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.766	46.100	46.100	2.766	46.100	46.100	2.138	35.635	35.635
2	1.304	21.739	67.839	1.304	21.739	67.839	1.932	32.204	67.839
3	.677	11.281	79.120						
4	.634	10.561	89.681						
5	.431	7.184	96.865						
6	.188	3.135	100.000						

Extraction Method: Principal Component Analysis.

Scree Plot



Component Matrix

	Component	
	1	2
X4	.722	.591
X6	.708	.605
X3	-.700	.314
X2	.694	-9.999E-02
X7	.666	-.321
X1	-.572	.614

Extraction Method: Principal Component Analysis.
a 2 components extracted.

Reproduced Correlations

		X1	X2	X	X4	X6	X
Reproduced Correlation	X1	.705	-.459	.59	-5.039E-02	-3.379E-02	-.57
	X2	-.459	.492	-.51	.442	.431	.49
	X3	.594	-.517	.58	-.320	-.306	-.56
	X4	-5.039E-02	.442	-.32	.870	.869	.29
	X6	-3.379E-02	.431	-.30	.869	.868	.27
	X7	-.579	.495	-.56	.291	.277	.54
Residual	X1		.160	-.11	-7.475E-02	-7.392E-02	.12
	X2	.160		8.517E-0	-8.672E-02	-9.513E-02	-.10
	X3	-.118	8.517E-02		-3.309E-02	6.387E-02	.21
	X4	-7.475E-02	-8.672E-02	-3.309E-0		-8.138E-02	-6.280E-0
	X6	-7.392E-02	-9.513E-02	6.387E-0	-8.138E-02		5.029E-0
	X7	.123	-.107	.21	-6.280E-02	5.029E-02	

Extraction Method: Principal Component Analysis.
a Residuals are computed between observed and reproduced correlations. There are 14 (93.0%) nonredundant residuals with absolute values > 0.05.
b Reproduced communalities

Rotated Component Matrix

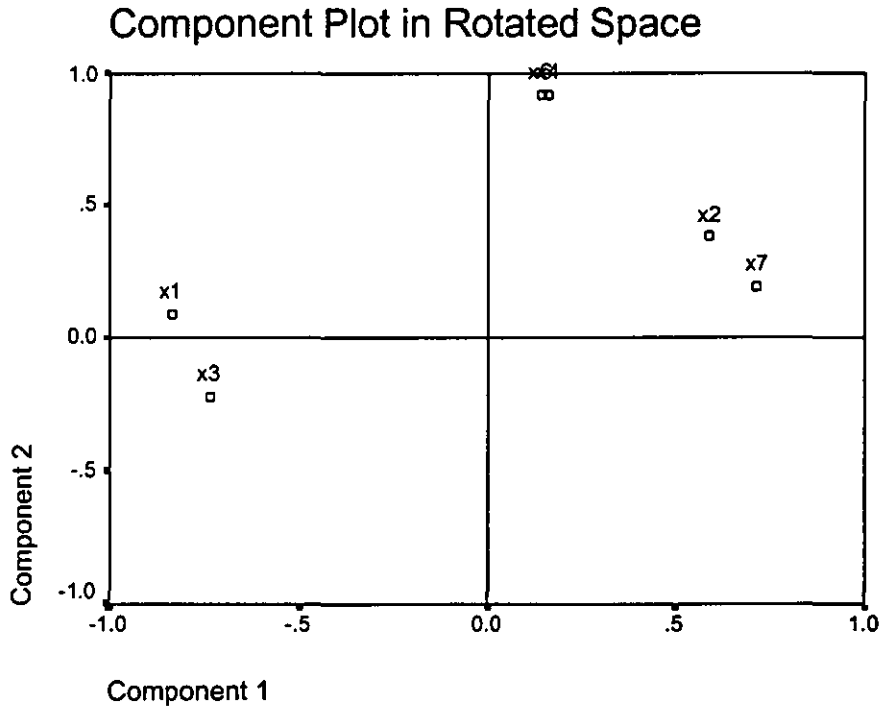
	Component	
	1	2
X1	-.835	8.868E-02
X3	-.735	-.221
X7	.714	.194
X2	.590	.379
X6	.138	.921
X4	.158	.919

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.
a Rotation converged in 3 iterations.

Component Transformation Matrix

Component	1	2
1	.755	.655
2	-.655	.755

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.



Component Score Coefficient Matrix

	Component	
	1	2
X1	-.465	.220
X2	.240	.107
X3	-.349	.016
X4	-.100	.513
X6	-.111	.518
X7	.343	-.028

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. Component Scores.

Component Score Covariance Matrix

Component	1	2
1	1.000	-1.338E-16
2	-1.338E-16	1.000

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. Component Scores.

Anexo 3. Base de Datos Pólizas de Seguros

Póliza	Edad	Antigüedad	Potencia	Grave	Clasificación	di
1	21	6	75	1	1	1.36929
2	48	10	130	1	1	1.37188
3	40	12	60	1	1	0.21466
4	28	5	90	1	1	1.08271
5	36	10	70	1	1	0.55472
6	24	1	75	1	1	0.35462
7	23	10	130	1	1	3.48902
8	38	2	75	1	2	-0.67886
9	23	6	75	1	1	1.19992
10	59	8	130	1	2	0.13608
11	24	1	60	1	2	-0.10373
12	23	10	110	1	1	2.87789
13	29	1	75	1	2	-0.06881
14	36	10	75	1	1	0.7075
15	24	1	70	1	1	0.20183
16	23	10	130	1	1	3.48902
17	25	5	60	2	1	0.42008
18	42	8	70	2	2	-0.25765
19	50	2	110	2	2	-0.62562
20	35	5	75	2	2	0.03157
21	33	2	60	2	2	-0.71378
22	19	3	70	2	1	0.92951
23	59	15	60	2	2	-0.938
24	63	12	70	2	2	-1.42755
25	65	5	110	2	2	-1.43953
26	46	5	60	2	2	-1.35833
27	38	1	60	2	2	-1.28934
28	42	4	60	2	2	-1.17171
29	29	1	75	2	2	-0.06881
30	22	2	90	2	1	1.13446
31	39	1	70	2	2	-1.06846
32	65	5	110	2	2	-1.43953
33	46	5	75	2	2	-0.89998
34	38	2	90	2	2	-0.22052
35	38	1	60	2	2	-1.28934
36	42	4	60	2	2	-1.17171
37	28	5	60	2	2	0.16602
38	59	15	70	2	2	-0.63244
39	63	12	70	2	2	-1.42755
40	65	5	110	2	2	-1.43953

Anexo 4. Base de Datos Pólizas de Seguros

Póliza	Edad	Antigüedad	Potencia	Grave	Clasificación	di
1	21	6	75	1	1	1.36929
2	48	10	130	1	1	1.37188
3	40	12	60	1	1	0.21466
4	28	5	90	1	1	1.08271
5	36	10	70	1	1	0.55472
6	24	1	75	1	1	0.35462
7	23	10	130	1	1	3.48902
8	38	2	75	1	2	-0.67886
9	23	6	75	1	1	1.19992
10	59	8	130	1	2	0.13608
11	24	1	60	1	2	-0.10373
12	23	10	110	1	1	2.87789
13	29	1	75	1	2	-0.06881
14	36	10	75	1	1	0.7075
15	24	1	70	1	1	0.20183
16	23	10	130	1	1	3.48902
17	25	5	60	2	1	0.42008
18	42	8	70	2	2	-0.25765
19	50	2	110	2	2	-0.62562
20	35	5	75	2	2	0.03157
21	33	2	60	2	2	-0.71378
22	19	3	70	2	1	0.92951
23	59	15	60	2	2	-0.938
24	63	12	70	2	2	-1.42755
25	65	5	110	2	2	-1.43953
26	46	5	60	2	2	-1.35833
27	38	1	60	2	2	-1.28934
28	42	4	60	2	2	-1.17171
29	29	1	75	2	2	-0.06881
30	22	2	90	2	1	1.13446
31	39	1	70	2	2	-1.06846
32	65	5	110	2	2	-1.43953
33	46	5	75	2	2	-0.89998
34	38	2	90	2	2	-0.22052
35	38	1	60	2	2	-1.28934
36	42	4	60	2	2	-1.17171
37	28	5	60	2	2	0.16602
38	59	15	70	2	2	-0.63244
39	63	12	70	2	2	-1.42755
40	65	5	110	2	2	-1.43953

Anexo 5. Base de datos para análisis de grupos

Provincia	PIB.PC	TASANAT	E.AGRI	E.IND	E.SERV
Las Tunas	1,673,914	9.7	19.95	32.61	47.44
Ciego de Ávila	1,349,165	9.2	21.82	34.81	43.37
Santiago de Cuba	2,135,255	11.9	10.16	21.09	68.75
Cienfuegos	1,213,720	11.8	28.67	28.44	42.89
Pinar del Río	1,361,496	10.0	28.32	27.56	44.12
Matanzas	1,426,991	11.0	22.89	33.13	43.98
La Habana	2,005,302	10.5	6.27	40.79	52.94
Holguín	1,264,544	9.9	43.83	19.81	36.36
Ciudad Habana	2,023,265	11.9	1.26	27.19	71.56
Camaguey	1,261,502	11.4	22.22	27.78	50.00
Santi Spíritus	1,806,340	10.2	13.02	42.01	44.97
Guantánamo	1,673,557	9.7	5.91	47.12	46.97
Villa Clara	1,690,932	10.6	16.67	41.67	41.67
Granma	1,619,684	11.8	7.74	33.64	58.62

Anexo 6. Validación del análisis de Grupos.

Quick Cluster

Initial Cluster Centers

	Cluster		
	1	2	3
PIB.PC	1.673.914	1.213.720	2.135.255
TASANAT	9.7	11.8	11.9
P_OCUPAGR	19.95	28.67	10.16
P_OCUPIND	32.61	28.44	21.09
P_OCUPSERV	47.44	42.89	68.75

Iteration History

	Change i Cluste Center		
Iteration		2	3
1	18971.40	99183.000	80647.667
2	.00	.000	.000

a Convergence achieved due to no or small distance change. The maximum distance by which any center has changed is .000. The current iteration is 2. The minimum distance between initial centers is 460194.000.

Cluster Membership

Case Number	PROVINC	Cluster	Distance
1	Las Tunas	1	18971.403
2	Ciego de Avila	2	36262.001
3	Santiago	3	80647.667
4	Cienfuegos	2	99183.000
5	P.Río	2	48593.000
6	Matanzas	2	114088.000
7	La Habana	3	49305.336
8	Holguín	2	48359.004
9	C.Habana	3	31342.335
10	Camaguey	2	51401.001
11	Santi Spíritus	1	113454.600
12	Guantánamo	1	19328.403
13	Villa Clara	1	1953.415
14	Granma	1	73201.401

Final Cluster Centers

	Cluster		
	1	2	3
PIB.PC	1.692.885	1.312.903	2.054.607
TASANAT	10.4	10.5	11.4
P_OCUPAGR	12.66	27.96	5.90
P_OCUPIND	39.41	28.59	29.69
P_OCUPSERV	47.93	43.45	64.42

Distances between Final Cluster Centers

Cluster	1	2	3
1		379982.400	361721.934
2	379982.400		741704.334
3	361721.934	741704.334	

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
PIB.PC	578446278951.672	2	5490483559.988	11	105.354	.000
TASANAT	1.098	2	.851	11	1.290	.314
P_OCUPAGR	588.202	2	47.980	11	12.259	.002
P_OCUPIND	176.867	2	44.763	11	3.951	.051
P_OCUPSERV	449.563	2	41.739	11	10.771	.003

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of Cases in each Cluster

Cluster		5.000
		6.000
		3.000
Valid		14.000
Missing		.000