ESCUELA SUPERIOR POLITECNICA DEL LITORAL

Instituto de Ciencias Matemáticas

"ANÁLISIS MULTIVARIADO DE LAS VENTAS DE UNA EMPRESA DE ALIMENTOS MARZO 1994-SEPTIEMBRE 2006"

TESIS DE GRADO

Previa a la obtención del Título de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

Presentado por:

JOSE M. AGUAYO ESCANDON

Guayaquil – Ecuador

2007

A G R A D E C I M I E N T O

A Dios por darme el camino y la fuerza para recorrerlo y terminarlo.

A mi familia por apoyar cada paso que daba por pequeño que fuera.

A mi esposa por el amor y valor que me dio para construir mi carrera y alcanzar esta meta.

Al Mat. John Ramírez, Director de Tesis, por su guía y paciencia para la elaboración de este estudio.

A mis jefes y compañeros de trabajo, por apoyar mí desarrollo profesional.

D E D I C A T O R I A

MIS PADRES

MIS HERMANOS

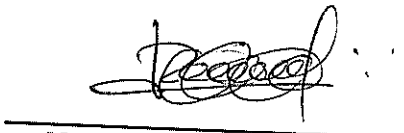
MI ESPOSA

MI HIJA

MIS AMIGOS

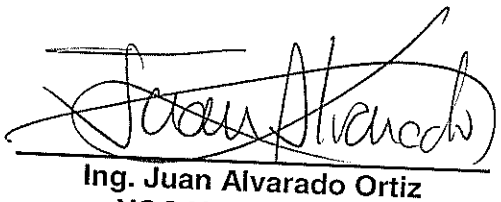
TRIBUNAL DE GRADO

Ing. Robert Toledo Echeverría
Presidente del Tribunal



Mat. John Ramirez Figueroa
DIRECTOR DE TESIS

Mat. Eduardo Rivadeneira Molina
VOCAL PRINCIPAL

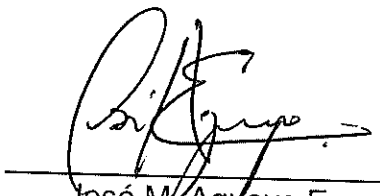


Ing. Juan Alvarado Ortiz
VOCAL PRINCIPAL

DECLARACIÓN EXPRESA

"La responsabilidad del contenido de esta Tesis de Grado, me corresponden exclusivamente: y el patrimonio intelectual de la misma a la ESCUELA SUPERIOR POLITECNICA DEL LITORAL."

(Reglamento de Graduación de la ESPOL)



José M. Aguayo Escandón

RESUMEN

El presente estudio nos plantea dos herramientas que una empresa de alimentos debería usar para solucionar problemas como el mal cálculo de presupuesto y mal enfoque del segmento al cual dirigir sus productos.

Utilizando el registro de compra en dólares de los clientes de la empresa estudiada se armaron tablas con registros desde marzo 1994. Mediante un Análisis de Series de Tiempo se obtuvieron pronósticos que se acercaron mucho a la venta real (en dólares).

En la segunda etapa se tomó una muestra de los clientes de Guayaquil atendidos por la empresa. Con esta información se realizó un Análisis de Componentes Principales y se obtuvo información muy importante er
a las características más importantes de compra de los consumidores

La aplicación de este estudio en una empresa de alimentos puede optimizar recursos y mejorar el enfoque comercial.

INDICE GENERAL

ABREVIATURAS

ACP	**Análisis de Componentes Principales**
AST	**Análisis de Serie de Tiempo**
FAS	**Función de Autocorrelación Simple**
FAP	**Función de Autocorrelación Parcial**

SIMBOLOGIA

MA(q)	**Modelo de Medias Móviles**
AR(p)	**Modelo Autorregresivo**
ARMA(p,q)	**Modelo de Autorregresivo y de Medias Móviles**
ARIMA(p,d,q)	**Modelo de Autorregresivo y de Medias Móviles**
X_i	**Variable de Investigación**
Y_i	**Variable Original para el ACP**

INDICE DE TABLAS

INDICE DE GRÁFICOS

Pág..

CAPITULO 2 Teoría Estadística

CAPITULO 3 Análisis Estadístico

CAPITULO I

1.1. ANTECEDENTES

La empresa, de la cual se omitirá el nombre por seguridad empresarial y políticas internas, tiene como objetivo producir, comercializar y distribuir productos cárnicos, conservas y arroz.

La empresa ha dividido la comercialización de sus productos en tres canales principales

- Autoservicios: Supermercados de gran tamaño con un número ilimitado de productos de distintas categorías con más de 4 cajas registradoras, en la cuales los consumidores acceden a los productos directamente. Este canal se divide en los siguientes Sub – Canales:
  - Autoservicios Mayores: con clientes con mas de 6 cajas registradoras.

- Autoservicios Menores: con clientes como Comisariatos/Supermercados, Otras Cadenas de Supermercados, Almacenes, Supermercados Independientes

❖ Tradicional: Agrupa a todos aquellos locales en los cuales el consumidor final se abastece de víveres, cárnicos, confitería, bebidas y artículos para consumo principalmente en el hogar. La ocasión de compra del consumidor en estos locales es el re - abastecimiento de cantidades limitadas de víveres por visita. Este Canal se divide en los siguientes Sub – Canales:

- Víveres: el cual comprende clientes como: Frigoríficos/Centros Cárnicos, Kioscos (fijos de víveres), Mayoristas, Mercados Populares, Tiendas Tradicionales, y Micromercados.
- Tiendas Especializadas: el cual posee clientes como Delicatessens, Estaciones de Servicio, Panaderías, Carnicería /Tercena, Farmacias y Licorerías, Otras Tiendas y Servicios Especiales

❖ Consumo Inmediato: Agrupa a los negocios de venta de comidas preparadas y bebidas. El objetivo es ofrecer al consumidor final servicio

de alimentación para consumir en el lugar, llevar o entregar a domicilio. Este Canal se divide en los siguientes Sub – Canales:

- Comidas y Bebidas: dentro de los cuales se encuentran clientes como restaurantes, asaderos, cafeterías/bares, clubes/resorts, fritaderías, hornaderías, comidas rápidas, pizzerías, y comedores populares
- Educación y Entretenimiento: donde se encuentran clientes como cines, clubes 4y 5 estrellas, bares y comedores de colegio y universidades, centros de esparcimientos o parques y centro de eventos.
- Hotelería y Transporte: en esta categoría se encuentran clientes como hoteles, hosterías, hostales y aerolíneas.
- Trabajo e Instituciones: donde se encuentran clientes como comedores de industrias, empresas, hospitales y clínicas, y servicios de catering.

Todos los canales difieren tanto en concepto como fuerza de venta. La fuerza de venta es especializada según el canal con un portafolio de productos muy diferente. Es decir podemos tener el mismo producto en

varios canales pero en diferentes presentaciones. Pongamos de ejemplos los embutidos. En el canal Tradicional y Autoservicios encontraremos paquetes de salchichas de 200 gramos pero estos no serían los ideales para el canal de Consumo Inmediato, por esta razón se comercializan las presentaciones de kilogramos.

Al tener una fuerza de venta para cada tipo de canal de comercialización se cuenta con un personal más especializado para comercializar los productos que describiremos a continuación.

Existen tres líneas básicas de productos en la empresa en cuestión tenemos: La línea de productos cárnicos, conservas y arroz.

- Entre los productos cárnicos que se comercializan en esta empresa se encuentran:
  - Cerdo: El cerdo es faenado, procesado y conservado bajo rigurosas políticas de calidad. Debido al bajo contenido en grasa es considerado el mejor cerdo del mercado. Entre la variedad de productos de cerdo tenemos: Canales, es decir cerdos enteros o mitades, cortes primarios; por ejemplos piernas, brazos, lomo

etcétera. También se producen empaques individuales, como por ejemplo de fritada, chuletas, patas etc. Y por último tenemos cortes al granel.

- Embutidos: Pollo y Res. De ambos se produce una amplia variedad productos embutidos: mortadela, salchichas (cóctel, hot dog, desayuno etc.), jamones. Y productos horneados. Entre estos últimos encontramos productos como pavo, fritada, y piernas de cerdo. Todos estos listos para hornear sin mayor esfuerzo.
- Productos Congelados: Esta línea se caracteriza por tener productos de fácil consumo, enfocado en consumidores sin mucho tiempo de cocinar después del trabajo. Dentro de las categorías en esta línea encontramos hamburguesas, apanados, marinados y otros
- Productos del Mar: Los mariscos también se destacan dentro del grupo objetivo del mercado mencionado en los productos congelados. Esta al tener presentaciones más económicas ingresa muy bien en el Canal Tradicional. En la variedad de productos en esta línea encontramos: Tilapia, camarón, pargo, atún, corvina, picudo, dorado y salmón

- La línea de conservas fue una línea netamente del Canal Autoservicios ha tomado mucha fuerza dentro del canal Tradicional ganando espacio en el mercado muy rápidamente. Entre los productos de conservas tenemos:
  - Aderezos: Salsas de tomate, mayonesa, mostaza, spaghetti y otros
  - Ajíes: Ajo , habanero, criollo y tabasco
  - Aliños: Salsas chinas, tabasco y pastas de tomate.
  - Culinarios: Aceite, cremas, sopas y caldos
  - Especialidades de dulce: Cóctel de frutas, duraznos en almíbar y mermeladas
  - Comida rápida: Enlatados como el maíz dulce, fréjol y lentejas.
- La línea de arroz se la comercializa empacada y esta dirigido al igual que como se vio en el ejemplo del embutido, con diferentes presentaciones según el canal.

Con esta pequeña reseña[1] se describirá cual es el problema que se plantea en esta tesis y cuales son los objetivos que ayudará a encontrar la solución para esta.

[1] Información de Estructura Comercial de la empresa a estudiar.

1.2. OBJETIVO DE LA TESIS

Se desea mediante este estudio otorgar a empresas, como la que se presentó, herramientas que le permitan encontrar:

1. Obtener una proyección o pronósticos de ventas más exactas que los lleve a tomar decisiones más acertadas en períodos futuros.

2. Obtener los factores que mejor explican la decisión de compra en cada línea.

1.3. METODOLOGIA

1. Utilizaremos modelos de series de tiempo para realizar pronósticos confiables que servirán para tomar decisiones más exactas.
2. Mediante una encuesta a los minoristas o clientes de esta empresa se desarrollara una investigación mercado para obtener la información correspondiente a los componentes principales que afectan la decisión de compra de las diferentes líneas y la relación que existe entre estos factores y los segmentos del mercado en el Canal Tradicional.

1.4. DESCRIPCION DE LAS VARIABLES

En esta investigación se tienen seis variables cuantitativas. Estas, en breves rasgos describen la venta en dólares de cada línea. A continuación tenemos la descripción detallada de cada una.

X_1: CERDO; indica las ventas en dólares de la línea de producto cárnicos derivados del faenamiento y procesamiento de cerdo.

X_2: EMBUTIDOS; indica las ventas en dólares de la línea de productos cárnicos procesados para la producción de embutidos.

X_3: CONGELADOS; indica la venta en dólares de la línea de productos cárnicos procesados de tal forma que sean productos listos para consumir luego de calentar.

X_4: PRODUCTOS DEL MAR; indica la venta en dólares de la línea de mariscos procesados.

X_5: CONSERVAS; indica la venta en dólares de la línea de productos secos procesados para convertirse en productos de conservas.

X_6: ARROZ; indica las ventas en dólares de la línea de arroz enfundado.

Para la investigación de mercado es necesario, antes de declarar las variables, plantear las hipótesis que surgen del problema de investigación. Este problema presenta el siguiente planteamiento:

1.5. DEFINICIÓN DEL PROBLEMA

1.5.1. DEFINICIÓN DEL PROBLEMA DE INVESTIGACIÓN DE MERCADO

No se dispone de información, complementaría a las ventas registradas desde marzo de 1994 hasta septiembre de 2006, que muestre los factores que afectan la decisión de compra.

Se desea obtener información que complemente el análisis de series de tiempo con relación a los factores que influyen en el proceso de compra de los consumidores en cada línea.

1.5.2. MERCADO OBJETIVO

Para análisis de componentes principales se tomara como mercado objetivo a los dueños de los negocios que pertenecen al canal tradicional.

1.5.3. HIPÓTESIS DE LA INVESTIGACIÓN DE MERCADO

H_{01}: El segmento femenino es el que mayor compra de cerdo realiza en el Canal Tradicional.

H_{02}: El segmento femenino es el que mayor compra de embutido realiza en el Canal Tradicional.

H_{03}: El segmento femenino es el que mayor compra de congelados realiza en el Canal Tradicional.

H_{04}: El segmento femenino es el que mayor compra de productos del mar realiza en el Canal Tradicional.

H_{05}: El segmento femenino es el que mayor compra de conservas realiza en el Canal Tradicional.

H_{06}: El segmento femenino es el que mayor compra de arroz realiza en el Canal Tradicional.

H_{07}: La mayoría de los compradores de cerdo son mayores a 30 años.

H_{08}: La mayoría de los compradores de embutidos son mayores a 30 años.

H_{09}: La mayoría de los compradores de congelados son mayores a 30 años.

H_{010}: La mayoría de los compradores de productos de mar son mayores a 30 años.

H_{011}: La mayoría de los compradores de conservas son mayores a 30 años.

H_{012}: La mayoría de los compradores de arroz son mayores a 30 años.

H_{013}: Existen una alta correlación entre todos los factores que afectan la decisión de compra de cerdo.

H_{014}: Existen una alta correlación entre todos los factores que afectan la decisión de compra de embutidos.

H_{015}: Existen una alta correlación entre todos los factores que afectan la decisión de compra de productos congelados.

H_{016}: Existen una alta correlación entre todos los factores que afectan la decisión de compra de productos del mar.

H_{017}: Existen una alta correlación entre todos los factores que afectan la decisión de compra de conservas.

H_{018}: Existen una alta correlación entre todos los factores que afectan la decisión de compra de arroz.

1.5.4. CUESTIONARIO

Con el objeto de analizar por segmento del mercado cada factor que influye en el proceso de compra de las líneas se ha desarrollado el siguiente cuestionario.

Analisis Multivariado de las ventas de una Empresa de Alimentos

Cuestionario

CANAL TRADICIONAL

2. De cada 10 compradores, Cuantos son hombres y cuantos son mujeres?

	Mujeres	**Hombres**
CERDO		
EMBUTIDOS		
CONGELADOS		
PRODUCTOS DEL MAR		
CONSERVAS		
ARROZ		

3. Determine con una **X** en que rango de edad se encuentran sus compradores de cada linea

	Menores a 15 anios	**Entre 15 y 30 anios**	**Mayor a 30 anios**
CERDO			
EMBUTIDOS			
CONGELADOS			
PRODUCTOS DEL MAR			
CONSERVAS			
ARROZ			

4. Califique con un **circulo** el grado de importancia de las siguientes caracteristicas que influyen en la decision de compra en cada linea.

	Nada Importante				**Muy Importante**
CERDO					
Calidad	1	2	3	4	5
Sabor	1	2	3	4	5
Textura	1	2	3	4	5
Empaque	1	2	3	4	5
Costumbre	1	2	3	4	5
Frescura	1	2	3	4	5
Precio	1	2	3	4	5
Promociones	1	2	3	4	5
Impulso	1	2	3	4	5
Publicidad	1	2	3	4	5
Otros	1	2	3	4	5
EMBUTIDOS					
Calidad	1	2	3	4	5
Sabor	1	2	3	4	5
Textura	1	2	3	4	5
Empaque	1	2	3	4	5
Costumbre	1	2	3	4	5
Frescura	1	2	3	4	5
Precio	1	2	3	4	5
Promociones	1	2	3	4	5
Impulso	1	2	3	4	5
Publicidad	1	2	3	4	5
Otros	1	2	3	4	5
CONGELADOS					
Calidad	1	2	3	4	5
Sabor	1	2	3	4	5
Textura	1	2	3	4	5
Empaque	1	2	3	4	5
Costumbre	1	2	3	4	5
Frescura	1	2	3	4	5
Precio	1	2	3	4	5
Promociones	1	2	3	4	5
Impulso	1	2	3	4	5
Publicidad	1	2	3	4	5
Otros	1	2	3	4	5
PRODUCTOS DEL MAR					
Calidad	1	2	3	4	5
Sabor	1	2	3	4	5
Textura	1	2	3	4	5
Empaque	1	2	3	4	5
Costumbre	1	2	3	4	5
Frescura	1	2	3	4	5
Precio	1	2	3	4	5
Promociones	1	2	3	4	5
Impulso	1	2	3	4	5
Publicidad	1	2	3	4	5
Otros	1	2	3	4	5
CONSERVAS					
Calidad	1	2	3	4	5
Sabor	1	2	3	4	5
Textura	1	2	3	4	5
Empaque	1	2	3	4	5
Costumbre	1	2	3	4	5
Frescura	1	2	3	4	5
Precio	1	2	3	4	5
Promociones	1	2	3	4	5
Impulso	1	2	3	4	5
Publicidad	1	2	3	4	5
Otros	1	2	3	4	5
ARROZ					
Calidad	1	2	3	4	5
Sabor	1	2	3	4	5
Textura	1	2	3	4	5
Empaque	1	2	3	4	5
Costumbre	1	2	3	4	5
Frescura	1	2	3	4	5
Precio	1	2	3	4	5
Promociones	1	2	3	4	5
Impulso	1	2	3	4	5
Publicidad	1	2	3	4	5
Otros	1	2	3	4	5

CAPÍTULO II

2. TEORIA ESTADISTICA

Para el desarrollo de la presente tesis se utilizaran cuatro técnicas de estadística aparte del análisis descriptivo de las diferentes variables.

1. Análisis de Series de Tiempo[2]

2. Análisis de Componentes Principales[3]

3. Muestreo Aleatorio Simple[4]

2.1 ANALISIS DE SERIES DE TIEMPO

Se plantea utilizar el Análisis de Series de Tiempo para, en función de las ventas registradas desde marzo de 1994 hasta septiembre de 2006, realizar estimaciones de un modelo que explique su comportamiento y obtener pronósticos de las ventas en los meses posteriores

[2] Para la teoría concerniente al Análisis de Series de Tiempo se utilizó el capítulo de de Modelos de Series de Tiempo disponible en el texto Econometría 2da. Edición 2000 por A. Novales, Mc Grawhill.
[3] En cuanto a la teoría utilizada para el Análisis de Componentes Principales se utilizo el capítulo 2 de Análisis de Componentes Principales diseñado por el Profesor Salvador Carrasco Arroyo de la Universidad de Valencia.
[4] La teoría de la etapa de muestreo fue tomada del curso dado por el ICM en Ingeniería en Estadística Informática.

Para la estimación del mejor modelo y sus correspondientes pruebas utilizaremos Demetra[5], y lo haremos en base de los modelo ARIMA. A continuación se encontrara la teoría concerniente a los modelos ARIMA.

2.1.1. INTRODUCCION

Una lista de variables exógenas se utiliza para explicar el comportamiento de otra variable, endógena. Hemos tratado del contraste de hipótesis con dichos modelos, así como de su utilización con fines predictivos y, posteriormente, hemos considerado modelos dinámicos. Este desarrollo convencional del modelo econométrico de relación entre variables se basa en el principio de mínimos cuadrados, tradicional en la Teoría Estadística, que se ha aplicado habitualmente en la estimación de relaciones en todo tipo de ciencias, experimentales o no experimentales.

Pero en ciencias no experimentales como la Economía, el investigador debe incluso tratar de descubrir la especificación del modelo, pues no existe un diseño experimental que haya generado los datos de que dispone. Apenas hemos tratado hasta ahora de la especificación de dichos modelos, excepto por algunas referencias conceptuales, nunca suficientemente precisas, al

[5] Demetra Versión 2.0, desarrollado para Eurostat por Jens Dosse y Servais Hoffmann.

aspecto de Teoría Económica en debate. En 1970, Box y Jenkins propusieron una metodología rigurosa para la identificación, estimación y diagnóstico de modelos dinámicos para datos de series temporales que, justificadamente, se ha convertido en una herramienta habitual en el análisis de series económicas, y que presentamos en este capítulo. En la primera parte analizamos modelos en los que el comportamiento de una variable se explica utilizando sólo su propio pasado, por lo que se conocen como modelos univariantes. Posteriormente, consideramos modelos dinámicos con variables explicativas, que se conocen como modelos de función de transferencia.

Conceptualmente, los modelos de función de transferencia no son distintos de los modelos econométricos que hasta ahora hemos considerado, con la diferencia de que los procedimientos de especificación y diagnóstico que se introducirán permitan que las propiedades dinámicas de las relaciones entre variables, así como la estructura estocástica del término de error, queden perfectamente recogidas en los modelos estimados.

Los modelos univariantes, por otra parte, son un complemento perfecto al estudio de relaciones entre variables. Un modelo univariante permite caracterizar adecuadamente muchas de las características de una variable económica en su dimensión temporal, y ello es importante, al menos por dos

razones:

a) Porque el investigador querrá que las variables que aparecen en ambos lados de un modelo de relación tengan en común algunas de sus características más importantes, como indicativo de que la relación que ha especificado es realmente relevante.

b) Porque constituye una referencia, relativamente sencilla de obtener, con la que comparar posibles modelos de relación que posteriormente pudieran estimarse.

Se definirá y caracterizará una amplia familia de estructuras estocásticas lineales, así como una familia de estadísticos que nos permitan escoger una de tales especificaciones como la más adecuada para representar la estructura estocástica de la variable económica que se está analizando.

2.1.2.DEFINICIONES

2.1.2.a. Proceso estocástico, ruido blanco, paseo aleatorio

Comenzamos introduciendo algunos conceptos fundamentales sobre los que se sustenta la teoría en que se desarrolla el siguiente marco teórico:

Definición 2.1. Se define al proceso estocástico como una sucesión de variables aleatorias

Aunque el índice que describe la sucesión de variables aleatorias que configura un proceso estocástico no necesita tener una interpretación concreta. La utilización de este concepto de Econometría confiere al índice t una interpretación como del período al que corresponde la variable aleatoria y_t. Esta definición es muy general, y las variables aleatorias y_t no precisan satisfacer ninguna propiedad en particular. Podrían carecer de algunos momentos, como varianza o esperanza; incluso su distribución marginal podría no existir.

Según que las variables y_t satisfagan unas u otras propiedades, tenemos un proceso estocástico de uno u otro tipo, como los que comenzamos ya a presentar:

$$\{y_t\} \quad t = -\infty,....,-2,-1,0,1,2,....,\infty.$$

Definición 2.2. Se llama ruido blanco a una sucesión de variables aleatoria con esperanza cero, igual varianza, e independientes en el tiempo. En lo sucesivo, denotamos un ruido blanco, por $\{\varepsilon_t\}$.

Definición 2.3. Una caminata aleatoria es un proceso estocástico { y_t }

cuyas primeras diferencias forman un proceso de ruido blanco, es decir:

$$\nabla y_t = \varepsilon_t$$

O lo que es Lo mismo:

$$y_t = y_{t-1} + \varepsilon_T, \qquad t = -\infty,...,-2,-1,0,1,2,...,\infty$$

donde ε_t es un ruido blanco.

Una forma de abordar el análisis estadístico de una serie temporal económica consiste en considerar que la serie temporal correspondiente a una variable como el consumo agregado es la realización de un proceso estocástico. Bajo este punto de vista, cada dato de la serie es una realización es decir, una muestra de tamaño 1 del proceso estocástico.

2.1.2.b. Estacionariedad

Definición 2.4. Un proceso estocástico $\{y_t\}$ es estacionario en sentido estricto si para toda m-tupla (t_1, t_2, ..., t_m) y todo entero k el vector de variables $(y_{t_1}, y_{t_2},..., y_{t_m})$ tiene la misma distribución de probabilidad conjunta que el vector $(y_{t_1+k}, y_{t_2+k},..., y_{t_m+k})$

Se examinará esta definición en detalle: ¿Qué ocurre cuando m = 1? En tal

caso, el conjunto de variables se reduce a una sola (y_t) y, de acuerdo con la definición, en un proceso estocástico estacionario este conjunto ha de tener la misma distribución independientemente del valor del índice t, es decir, que las variables aleatorias que componen un proceso estocástico estacionario están idénticamente distribuidas. En particular, la esperanza y la varianza de las variables y_t son independientes del tiempo.

Cuando m = 2, la distribución conjunta del par (y_t, y_{t-s}) en un proceso estacionario debe ser independiente del tiempo. Como consecuencia, la covarianza entre ellas, así como su coeficiente de correlación, debe ser invariante en t, dependiendo únicamente del valor del retardo, s. Es decir, en un proceso estocástico estacionario, la covarianza entre y_t e y_{t-2} es igual a la covarianza entre y_s e y_{s-2} para todo t y s, si bien será generalmente diferente de la covarianza entre y_t e y_{t-1} o de la covarianza entre y_t e y_{t-3} que serán a su vez constantes en el tiempo, aunque diferentes entre si.

Sin embargo, el concepto de estacionariedad en sentido estricto implica el cumplimiento de un número de condiciones que es excesivo para nuestras necesidades prácticas. Generalmente, nos conformamos con un concepto menos exigente, como es el de estacionariedad en sentido débil o de segundo orden, que se produce cuando todos los momentos de primer y

segundo orden del proceso estocástico son invariantes en el tiempo. Estos momentos incluyen la esperanza matemática y la varianza de las variables y y_t pero también las covarianzas y correlaciones entre diversos retardos a que antes hemos hecho referencia. En lo sucesivo, cuando hablamos de proceso estacionario nos referimos a un proceso estocástico estacionario de segundo orden.

Algunas de estas condiciones son fácilmente contrastables. Por ejemplo, si una serie temporal muestra una tendencia creciente, como cada observación es una realización de la variable aleatoria correspondiente, no podremos mantener que la esperanza matemática de dichas variables es constante, sino que deberemos aceptar que crece con el tiempo. En otras ocasiones, una, serie temporal económica experimenta fluctuaciones de amplitud creciente en el tiempo, en cuyo caso deberemos reconocer que la varianza de las variables y_t no es constante y el proceso no es estacionario de segundo orden. El ruido blanco es un ejemplo de proceso estocástico estacionario de segundo orden.

La caminata aleatoria no es un proceso estacionario, puesto que puede escribirse:

$$y_t = y_{t-1} + \varepsilon_t = y_{t-2} + \varepsilon_{t-1} + \varepsilon_t = y_{t-3} + \varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t + \cdots$$

sin que esté definida su esperanza ni su varianza, ni tampoco la distribución marginal de cada y_t. Sí está definida, sin embargo, su distribución condicional en el pasado: { $y_{t-1,}$ $y_{t-2,\ldots,}$}, que será del mismo tipo que la de ε_t , (es decir normal o binomial dependiendo de que ésta lo sea), con igual varianza, y esperanza igual a la de ε_t , aumentada en el valor realizado de

$$y_{t-1,}$$

A veces se considera una caminata aleatoria al que se le ha impuesto una condición inicial, la de comenzar a partir de un valor numérico y_0. de modo que puede escribirse:

$$y_t = y_{t-1} + \varepsilon_t = y_{t-2} + \varepsilon_{t-1} + \varepsilon_t = y_{t-3} + \varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t + \cdots =$$

$$= y_o + \varepsilon_t + \varepsilon_{t-1} + \cdots + \varepsilon_2 + \varepsilon_1$$

Este proceso, sobre el que hay que insistir que es distinto de la caminata aleatoria que antes hemos definido, tiene distribución marginal, que será, por ejemplo Normal si la de ε_t lo es para todo t**,** con esperanza igual a y_0 y varianza igual a $t\sigma^2{}_\varepsilon$ creciente en el tiempo. Como consecuencia, tampoco es estacionario. Su distribución condicional es igual a la de la caminata aleatoria.

Unos estadísticos fundamentales en la especificación de modelos univariantes son las funciones de autocovarianza, de autocorrelación simple y de

auto-correlación parcial que a continuación introducimos:

Definición 2.5. La función de autocovarianza de un proceso estocástico

$\{y_t\}$ es una función, a la que en lo sucesivo nos referimos como FAC, que para cada instante t y cada número entero k toma un valor, denotado por $\gamma_K(t)$ igual a la covarianza entre y_t e y_{t-k} es decir:

Definición 2.6. La $\gamma_k(t) = Cov(y_{t`}, y_{t-k})$ función de autocorrelación simple de un proceso estocástico { y_t }, a la que en lo sucesivo nos referimos por FAS., es una función que para cada instante t y cada entero k toma un valor $\rho_K(t)$ igual a la correlación entre y_t e y_{t-k}

$$\rho_k(t) = \frac{Cov(y_{t`}, y_{t-k})}{\sqrt{Var(y_y)}\sqrt{Var(y_{t-k})}} = \frac{\gamma_k(t)}{\sqrt{Var(y_{t-k})}}$$

Definición 2.7. La función de autocorrelación parcial de un proceso estocástico $\{y_t\}$,

a la que en lo sucesivo nos referimos como FAP, es una función que para cada instante t y cada entero k toma un valor igual a la correlación entre y_t e

y_{t-k} , ajustada por el efecto de los retardos intermedios $y_{t-1,}$ $y_{t-2,\ldots}$ y_{t-k-1} .

El gran interés de un proceso estocástico estacionario reside en que las FAC, FAS y FAP son independientes del tiempo t, por lo que puede omitirse dicho argumento temporal. Lo que es crucial es que dicha invarianza permite la estimación muestral de tales funciones. Asimismo, aun antes de pasar a su estimación, el concepto de estacionariedad nos permite deducir algunas propiedades de estas funciones:

En general, parte de la correlación existente entre y_t e y_{t-2} estará producida por el hecho de que ambas están correlacionadas con y_{t-1} y eso es lo que trata de corregir la FAP. El primer valor de la FAP es la correlación entre y_t e y_{t-1} sin que haya que corregir por ningún retardo intermedio, puesto que no existen. Por eso es que el primer valor de las FAS y FAP de cualquier proceso estocástico coinciden. De manera análoga el valor inicial de la FAS ρ_0 es igual a 1 en todo proceso estacionario, por ser el cociente de la varianza del proceso (constante en el tiempo) por sí misma. Por un razonamiento similar al anterior, concluimos que el valor inicial de la FAP es asimismo igual a 1 en todo proceso estacionario. Por último, las FAS y FAP de un proceso de ruido blanco son iguales a cero, excepto en sus valores iniciales que, como hemos visto, son iguales a 1. Esto no es sino una

manifestación de una propiedad más general que consiste en que las FAS y FAP de todo proceso estocástico estacionario decrecen rápidamente hacia cero.

2.1.2.c. Estimación de las funciones de autocorrelación de un proceso estacionario

Ya se ha expuesto cómo la estacionariedad de un proceso permite considerar la estimación de sus FAS y FAP que son las herramientas básicas de la especificación de su representación univariante. La estimación de los distintos valores ρ_k se lleva a cabo de la siguiente forma:

$$\rho_k = \frac{\frac{1}{T-k}\sum_{k+1}^{r}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sqrt{\frac{1}{T}\sum_{1}^{T}(y_t - \bar{y})^2}\sqrt{\frac{1}{T-k}\sum_{k+1}^{T}(y_{t-k} - \bar{y})^2}}$$

Distintas simplificaciones se consiguen en esta expresión cuando el tamaño maestral T es grande con respecto a k, pues entonces dividir por T o por T-k es prácticamente lo mismo, en cuyo caso todos los cocientes de la forma $\frac{1}{T}\, y \frac{1}{T-k}$ que aparecen en la expresión anterior pueden eliminarse.

Por otra parte, las medias muestrales de $\left(y_t - \bar{y}\right)^2$ sobre las observaciones

1,2, ... , T, o sobre las observaciones k +1, k+2 ,…, T serán muy similares- Con estas aproximaciones se tiene el estimador:

$$r_k = \frac{\sum_{k+1}^{t}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{1}^{t}(y_t - \bar{y})^2} \quad k = 0,1,2....$$

La expresión que acabamos de presentar para la estimación de esta función garantiza que el valor estimado de $\hat{\gamma}_0$ será siempre igual a 1.

Finalmente, en todo proceso estacionario, la función de autocovarianza es simétrica, es decir, $\gamma_{-k} = \gamma_k$, lo que proviene del hecho de que la covarianza entre y_t e y_{t-k} es igual a la covarianza entre y_t e y_{t+k}. Cómo consecuencia, la función de autocorrelación es también simétrica, por lo que en el trabajo práctico se analizan únicamente dichas funciones para valores de k= 0, 1, 2,…

El primer valor de la FAP, que vamos a denotar por ϕ_{11} (luego veremos por qué utilizamos dos subíndices), puede estimarse transformando la serie y_t, en desviaciones con respecto a su media muestral $\tilde{y}_t = y_t - \bar{y}$, y a continuación estimando una regresión de $\tilde{y}_t$ sobre $\tilde{y}_{t-1}$

En el caso de variables con media muestral igual a cero, el estimador MCO

de β de la regresión $y_t = \beta x_t + u_t$ es precisamente igual al coeficiente de correlación parcial entre x e y, multiplicado por el cociente de sus desviaciones típicas. En este caso, las variables en la regresión son $\tilde{y}_t$, e $\tilde{y}_{t-1}$ y si el proceso es estacionario, sus desviaciones típicas son iguales. Por tanto, la pendiente estimada de la regresión anterior coincide con el coeficiente de correlación entre , y_t e y_{t-1} es decir, ϕ_{11}.

Otra consecuencia de este comentario es que el primer valor de la FAP es siempre igual al primer valor de la FAS. Intuitivamente, la razón es que al estimar la correlación entre y_t e y_{t-1} no hay que corregirla de ningún valor intermedio, por lo que las dos funciones, ϕ_{11}y γ_1 toman el mismo valor numérico, al igual que ocurre con sus valores teóricos.

El segundo valor de la FAP se estima mediante una regresión $\tilde{y}_t$ de sobre $\tilde{y}_{t-1}$ e $\tilde{y}_{t-2}$.

$$\tilde{y}_t = \phi_{21}\tilde{y}_{t-1} + \phi_{21}\tilde{y}_{t-2} + u_t$$

El coeficiente estimado ϕ_{22} mide la correlación entre $\tilde{y}_t$ e $\tilde{y}_{t-2}$ una vez que se ha tenido en cuenta el efecto común de $\tilde{y}_{t-1}$, incluida como otra variable explicativa adicional. Así, la FAP puede estimarse mediante una serie de regresiones, cada una de las cuales contiene como variable explicativa un retardo más que la anterior, y de las que nos vamos quedando en cada caso

con los coeficientes estimados en los retardos más altos: ϕ_{11}, ϕ_{22}, ϕ_{33}, …

Otra posibilidad de obtener la FAP estimada es mediante fórmulas recursivas.

Utilizando la FAS previamente estimada, y utilizando las ecuaciones de Yule-Walker, como veremos en la sección siguiente.

2.1.3.MODELOS AUTORREGRESIVOS

Comenzamos a introducir en esta sección las estructuras estocásticas lineales que trataremos de asociar a una serie temporal de datos económicos. La primera de tales estructuras es la de los modelos autorregresivos.

Definición 2.8. Un proceso autorregresivo de orden 1, denotado por AR(1), viene definido por

$$y_t = \phi y_{t-1} + \delta + \varepsilon_t$$

donde ϕ y δ son constantes y ε_t, es un ruido blanco.

Si un modelo AR(1) es estacionario, entonces su esperanza y varianza son constantes en el tiempo, y se tiene que:

$a) Ey_t = \phi Ey_{t-1} + \delta + E\varepsilon_t = \phi Ey_{t-1} + \delta, pero Ey_t = Ey_{t-1}$ por lo que

$Ey_t = \mu_y = \delta/(1-\phi)$

b) $Var(y_t) = \phi^2 \quad Var\ y_{t-1} + Var\ \varepsilon_t$ ahora bien, si le proceso AR(1) es estacionario, entonces $Var\ y_t$ = Var y_{t-1} , por lo que $Var(y) = \sigma_s^2/(1-\phi^2)$

Por otra parte, mediante sustituciones repetidas, puede verse que:

$$y_t = \left(\delta + \delta\phi + \delta\phi^2 + \cdots\right) + \left(\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \cdots\right) = [\delta/(1-\phi)] + \sum_{s=0}^{\infty}\phi^s\varepsilon_{t-s}$$

Esta expresión tendrá sentido sólo si la suma infinita que en ella aparece converge. Dicha suma es aleatoria, puesto que es una combinación lineal de las variables aleatorias ε_t y converge si y sólo si lo hace su varianza. Dado que las variables ε_t, son independientes, se tiene:

$$Var(y_t) = Var\left(\sum_0^{\infty}\phi^s\varepsilon_{t-s}\right) = \sum_0^{\infty}Var\left(\phi^s\varepsilon_{t-s}\right) = \sum_0^{\infty}\phi^{2s}Var\,\varepsilon_{t-s} = \sum_0^{\infty}\phi^{2s}\sigma^2{}_{\varepsilon} = \phi_s^2/(1-\phi^2)$$

por lo que la varianza de la combinación lineal es finita si y sólo si $|\phi| < 1$. Sólo si $|\phi| < 1$ tendrá sentido la expresión y, en consecuencia, y_t podrá expresarse como función de ε_t y de todas las variables ε anteriores al instante t, pero de ninguna futura. Sólo entonces es el proceso AR(1)

estacionario, para lo que es necesario y suficiente que $|\phi| < 1$. Es entonces cuando su esperanza y varianza están bien definidas por las expresiones que aquí hemos derivado.

El coeficiente de cada variable ε_{t-k} en dicha expresión es ϕ^k. Como el proceso ε_t es un ruido blanco, entonces se concluye que:

a) $E(y_{t-k}\varepsilon_t) = 0$ para todo k> 0, ya que y_{t-k} depende de ε_{t-k} y valores anteriores del proceso ε_t, pero no de sus valores futuros. Por tanto:

$$E(y_{t-k}\varepsilon_t) = E[\delta\varepsilon_t/(1-\phi)] + \sum_{s=0}^{\infty}\phi^s E(\varepsilon_{t-k-s}\varepsilon_t) = 0 + 0 = 0, \forall k > 0$$

b) $$E(y_t\varepsilon_{t-k}) = E[\delta\varepsilon_{t-k}/(1-\phi)] + \sum_{s=0}^{\infty}\phi^s E(\varepsilon_{t-s}\varepsilon_{t-k}) = \phi^k\sigma_s^2, \forall k \geq 0$$

Por otra parte, si se sustituye δ por su expresión equivalente $(1-\phi)\mu_y$ en el modelo se tiene:

$$y_t - \mu_y = \phi(y_{t-1} - \mu_y) + \varepsilon_{s`} \text{ es decir } \tilde{y}_t = \phi\tilde{y}_{t-1} + \varepsilon_t$$

Donde $\tilde{y}_t$ denota la diferencia entre y_t y su esperanza poblacional. Dicha diferencia tiene los mismos momentos que la variable $\tilde{y}_t$. En particular: $\sigma_y^2 = \sigma_{\tilde{y}}^2$, y pueden aplicarse los resultados anteriores a) y b). Con todo ello, la función de autocovarianza de este proceso resulta ser:

$y_0 = \sigma_y^2 = \sigma_t^2 /(1-\phi^2)$, como vimos antes

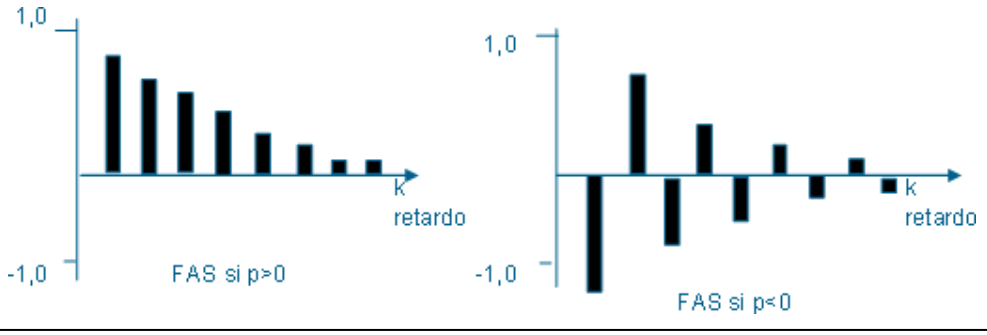
$$y_1 = E(\tilde{y}_t \tilde{y}_{t-1}) = E(\phi \tilde{y}_{t-1} + \varepsilon_t \tilde{y}_{t-1}) = \phi y_0$$

$$y_2 = E(\tilde{y}_t \tilde{y}_{t-2}) = E(\phi \tilde{y}_{t-1} \tilde{y}_{t-2} + \varepsilon_t \tilde{y}_{t-2}) = \phi y_1 + E(\varepsilon_t \tilde{y}_{y-2}) = \phi^2 y_0 + 0 = \phi^2 y_0$$

$$y_3 = E(\tilde{y}_t \tilde{y}_{t-3}) = E(\phi \tilde{y}_{t-1} \tilde{y}_{t-3} + \varepsilon_t \tilde{y}_{t-3}) = \phi y_2 + E(\varepsilon_t \tilde{y}_{t-3}) = \phi^3 y_0$$

Y así sucesivamente, de modo que: $p_0 = 1; p_1 = \phi; p_2 = \phi^2; ....; p_k = \phi^k$ para todo k> 0, por lo que los valores de la FAS son las sucesivas potencias del parámetro ϕ. Además, la condición $|\phi| < 1$ garantiza que los sucesivos valores de la FAS convergen a cero. Esta función puede tener dos aspectos distintos, dependiendo del signo de tal parámetro. Así, se tiene que:

Gráfico 2.1.1.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Función de Autocorrelación Simple



Autor: José Aguayo E. Fuente: Econometría 2da. Edición 2000 por A. Novales, Mc Grawhill

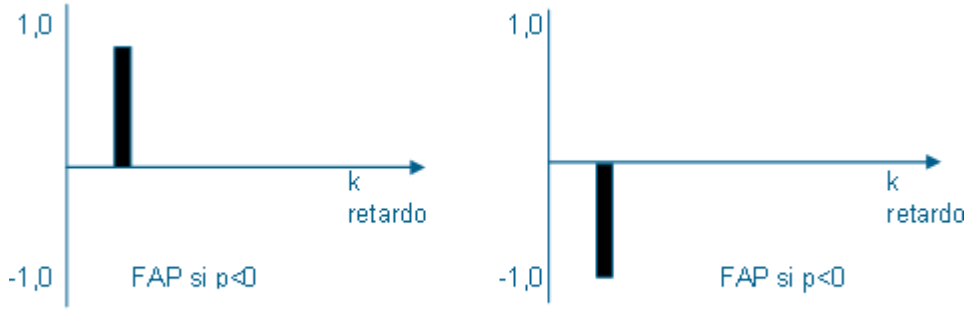
Se deduce que su FAP tiene como primer valor $\phi_{11} = \phi$ y como restantes valores $\phi_{kk} = 0$.

Por ejemplo, el segundo valor de dicha función sería el parámetro ϕ_{22} en la regresión:

$$\tilde{Y}_t = \phi_{21}\tilde{y}_{t-1} + \phi_{22}\tilde{y}_{t-2} + \varepsilon_t$$

pero, de acuerdo con el modelo teórico ϕ_{22} debe ser cero, y lo mismo ocurre para todo ϕ_{kk} con $k \geq 2$.

Grafico 2.1.2.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Función de Autocorrelación



Autor: José Aguayo E. Fuente: Econometría 2da. Edición 2000 por A. Novales, Mc Grawhill

Definición 2.9. Un proceso es autorregresivo de orden 2, que se denota como AR(2), si responde a la ley estocástica:

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$$

y es estacionario si $|\phi_2| < 1, \phi_1 + \phi_2 < 1$ y $\phi_2 - \phi_1 < 1$. Bajo estas condiciones, y_t, puede escribirse como función de ε_t , y sus valores previos. Calculemos la función de autocorrelación de un proceso AR(2) estacionario para ello comencemos notando que $Ey_t = \mu_y = \delta/(1-\phi_1-\phi_2)$. Ahora sustituimos δ en la ecuación del modelo por $(1-\phi_1-\phi_2)\mu_{y`}$, se tiene:

$$y_t - \mu_y = \phi_1(y_{t-1} - \mu_y) + \phi_2(y_{t-2} - \mu_y) + \varepsilon_t$$

Multiplicando por $y_{t-k} - \mu_y$ para k = 0, 1, 2 y tomando esperanzas se tiene que:

$$y_0 = \phi_1 y_1 + \phi_2 y_2 + \sigma_\varepsilon^2$$

$$y_1 = \phi_1 y_0 + \phi_2 y_1$$

$$y_2 = \phi_1 y_1 + \phi_2 y_0 \qquad [2.1.4]$$

y, en general, multiplicando $y_t - \mu_y = \phi_1(y_{t-1} - \mu_y) + \phi_2(y_{t-2} - \mu_y) + \varepsilon_t$ por $y_{t-k} - \mu_y$ para algún k>O y tomando esperanzas se tiene:

$$y_k = \phi_1 y_{k-1} + \phi_2 y_{k-2} \quad k \geq 1$$

donde hemos utilizado el hecho de que la función de autocovarianza es simétrica, de modo que $y_k = y_{-k}$ El sistema de ecuaciones puede resolverse para obtener:

$$Var(y_t) = y_0 = (1-\phi_2)\sigma_\varepsilon^2 /(1+\phi_2)\left[(1-\phi_2)^2 - \phi_1^2\right]$$

$$y_1 = \phi_1 y_0 /1-\phi_2$$

$$y_2 = \phi_2(1-\phi_2) + \phi_1^2 /1-\phi_2 * y_0$$

que implican:

$$\rho_1 = \phi_1 /(1-\phi_2)$$

$$\rho_2 = [\phi_1^2 /(1-\phi_2)] + \phi_2$$

y en general:

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2},\ k \geq 1$$

Partiendo de estimadores de $\sigma_y^2, \phi_1, \phi_2, \sigma_\varepsilon^2$ se pueden obtener estimaciones de $r_0.r_1, r_2$ o, recíprocamente, como es más usual en la práctica, podemos comenzar de los valores estimados en la muestra de $\sigma_y^2, \rho_0, \rho_1, \rho_2$ y obtener estimaciones de $\phi_1, \phi_2, \sigma_\varepsilon^2$. Esta estrategia dual se puede extender a modelos autorregresivos de orden p, AR(p). Para cualquier p>O. El sistema de ecuaciones [2.1.4] constituye las llamadas ecuaciones de Yule-Walker, que pueden extenderse a procesos AR(p)

Un proceso AR(2) con parámetro ϕ_2 negativo puede generar raíces complejas en su ecuación característica $1-\phi_1 L-\phi_2 L^2 = 0$. En tal caso, y_t, pre-

sentará ciclos de período T dado por.

$$\cos(2\pi / T) = \phi_1 /(2\sqrt{-\phi_2})$$

y factor de amortiguamiento: $d = \sqrt{-\phi_2}$ -

La función de autocorrelación de los procesos AR(2) también converge exponencialmente a cero. Sin embargo, dicha convergencia puede presentar una gran variedad de aspectos. Puede ser siempre con signo positivo, pero también puede alternar en signo. Estas dos eran las posibilidades en los modelos AR(1). Sin embargo, ahora puede haber también convergencia a cero siguiendo una curva sinusoidal. Esta diversidad, junto con el hecho de que para algunos valores de ϕ_1 y ϕ_2 la FAS es similar a la de los modelos AR(1) hace que sea preciso alguna herramienta adicional para identificar un proceso AR(2). En cuanto a su FAP, es claro que las estimaciones de los parámetros ϕ_{11} y ϕ_{22} en las regresiones.

$$\tilde{y}_t = \phi_{11}\tilde{y}_{t-1} + u_t$$

$$\tilde{y}_t = \varphi_{21}\tilde{y}_{t-1} + \phi_{22}\tilde{y}_{t-2} + \varepsilon_t$$

Serán no nulas, pero, sin embargo, la estimación de ϕ_{33} en

$$\tilde{y}_t = \phi_{31}\tilde{y}_{t-1} + \phi_{32}\tilde{y}_{t-2} + \varphi_{33}\tilde{y}_{t-3} + u_t$$

será estadísticamente igual a cero, al igual que ϕ_{44}, ϕ_{55}.... Esto implica que la FAP de un proceso AR(2) es cero para valores k> 2. Lo importante es que a pesar de la variedad de formas que puede adoptar la FAS dependiendo de los valores de los parámetros ϕ_1 y ϕ_2, sin embargo, la propiedad que acabamos de mencionar para la FAP es independiente de dichos valores.

Una propiedad similar es válida para todo modelo AR(p), donde p es cualquier entero positivo p: su función de autocorrelación parcial es cero para valores k > p.

2.1.4.MODELOS DE MEDIAS MOVILES

Se analizará en esta sección una clase diferente de procesos estocásticos, los procesos de medias móviles.

Definición 2.10. Se llama proceso de medias móviles de orden 1, que denotamos MA(1), a la estructura

$$y_t = \delta + \varepsilon_t - \theta\varepsilon_{t-1}$$

donde ε_t, es un ruido blanco. Como primeras propiedades de este proceso se tiene inmediatamente de su definición que E y_t, **=** δ y también que

$$Var(y_t) = (1+\theta^2)\sigma_\varepsilon^2$$

En cuanto a la función de autocovarianza, se tiene:

$$y_0 = \sigma_y^2 = (1+\theta^2)\sigma_\varepsilon^2$$

$$y_1 = -\theta\sigma_\varepsilon^2$$

$$y_2 = 0$$

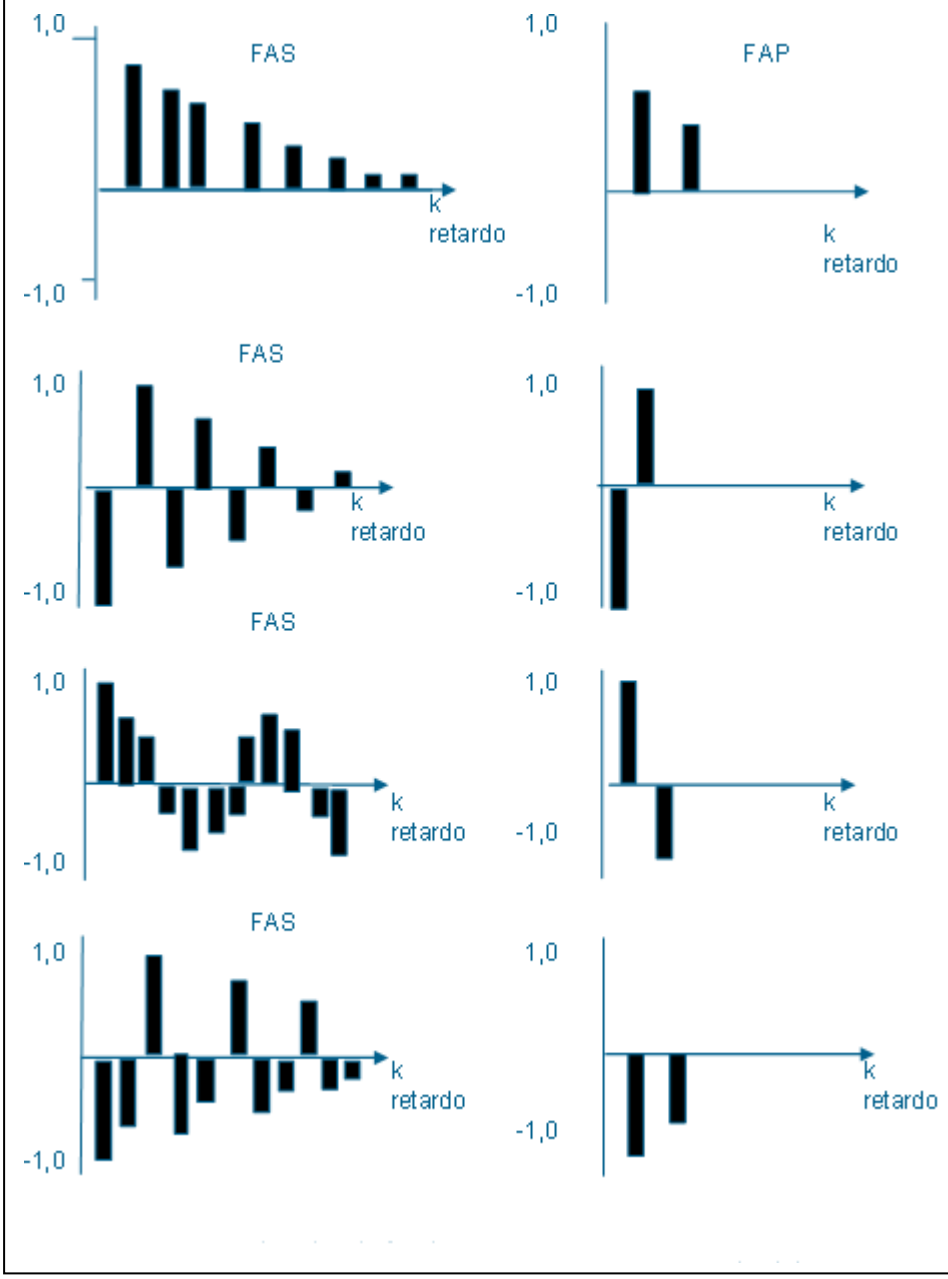
$$y_k = 0 \text{ para todo } k \geq 2$$

y por tanto: $\rho_1 = -[\theta/(1+\theta^2)]; \rho_2 = 0, \rho_k = 0$ para todo $k \geq 2$. La función $g(\theta) = -\theta/(1+\theta^2)$ es monótona decreciente en θ y, como consecuencia, puede verse fácilmente que el máximo valor absoluto que puede tomar ρ_1 en un modelo MA(1) es 0,50, y éste es el único valor no nulo de su función de autocorrelación simple, siendo negativo si θ > 0, y positivo en caso contrario.

Grafico 2.1.3.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

FAS vs FAP



Autor: José Aguayo E. Fuente: Econometría 2da. Edición 2000 por A. Novales, Mc Grawhill

A partir del proceso MA(1) puede llevarse a cabo el siguiente esquema «de inversión»:

$$Y_t = \delta + \varepsilon_t - \theta\varepsilon_{t-1} \text{ que implica: } \varepsilon_t = y_t - \delta + \theta\varepsilon_{t-1}$$

Del mismo modo:

$$\varepsilon_{t-1} = y_{t-1} - \delta + \theta\varepsilon_{t-2}$$

$$\varepsilon_{t-2} = y_{t-2} - \delta + \theta\varepsilon_{t-3}$$

por lo que, se tiene, sustituyendo:

$$\varepsilon_t = y_t - \delta + \theta(y_{t-1} - \delta + \theta\varepsilon_{t-2}) = y_t + \theta y_{t-1} - \delta(1+\theta) + \theta^2\varepsilon_{t-2} =$$

$$= y_t - \theta y_{t-1} - \delta(1+\theta) + \theta^2(y_{t-2} - \delta + \theta\varepsilon_{t-3}) =$$

$$= y_t + \theta y_{t-1} + \theta^2 y_{t-2} - \delta(1+\theta+\theta^2) + \theta^3\varepsilon_{t-3}$$

Por tanto:

$$y_t = -\theta y_{t-1} - \theta^2 y_{t-2} + \delta(1+\theta+\theta^2) - \theta^3\varepsilon_{t-3} + \varepsilon_t \qquad [2.1.5]$$

y el proceso continuaría, eliminando ahora el término ε_{t-3}.

Este procedimiento conduce, en el límite, a expresar y_t, como función de su propio pasado, así como del valor contemporáneo del ruido blanco ε_t, pero sólo tiene sentido si $|\theta| < 1$. De otro modo, se tendría en [2.1.5] que el pasado de y_t, tiene una gran importancia para determinar su comportamiento actual, y tanta más importancia cuanto más atrás en el tiempo. Cuando $|\theta| < 1$ se

dice que el proceso MA(1) es *invertible* y podemos obtener su representación autorregresiva como límite del procedimiento descrito en [2.1.5]:

$$y_t = \delta / 1 - \theta - \sum_{s=1}^{\infty} \theta^s y_{t-s} + \varepsilon_t$$

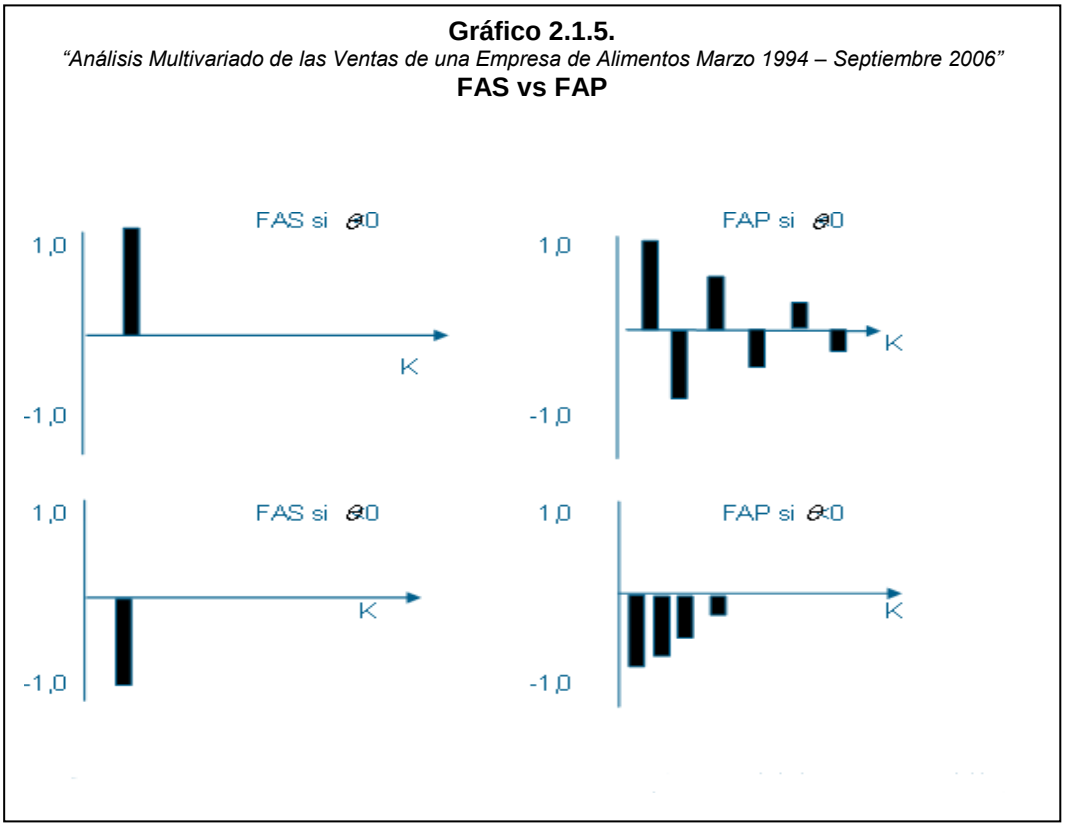
El hecho de que al invertir un proceso MA(1) se tengan coeficientes θ^s en los retardos de y_t, sugiere, como así es, que la FAP de un proceso MA(1) decae exponencialmente hacia cero, quizá alternando en signo.

Si el parámetro θ es negativo, entonces la FAP converge a cero exponencialmente alternando en signo, y empezando de un valor positivo. Si, en cambio, el parámetro θ tiene signo positivo, entonces la convergencia va a ser con todos los valores de la FAP. tomando signo negativo. Nótese, por tanto, que un proceso MA(1) no puede generar nunca una FAP que sea siempre positiva, pero sí puede generar una función de autocorrelación parcial que es siempre negativa.

Definición 2.11. Un proceso de medias móviles de orden 2 es un proceso estocástico que sigue la ley:

$$y_t = \delta + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2}$$

Gráfico 2.1.5.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
FAS vs FAP



Autor: José Aguayo E. Fuente: Econometría 2da. Edición 2000 por A. Novales, Mc Grawhill

Siguiendo un proceso de inversión análogo al que hicimos con el proceso MA(1) se puede probar fácilmente que la FAP de este proceso puede tener diversas formas, dependiendo de los signos y los valores relativos de parámetros θ_1 y θ_2 En cambio, la función de autocovarianza cumple:

$$y_0 = \sigma_y^2 = (1 + \theta_1^2 + \theta_2^2)\sigma_\varepsilon^2$$

$$y_1 = E[y_t y_{t-1}] = E[(\varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2})(\varepsilon_{t-1} - \theta_1\varepsilon_{t-2} - \theta_2\varepsilon_{t-3})] =$$

$$= E\varepsilon_t\varepsilon_{t-1} - \theta_1\varepsilon_t\varepsilon_{t-2} - \theta_2\varepsilon_t\varepsilon_{t-3} - \theta_1 E\varepsilon_{t-1}^2 + \theta_1^2 E\varepsilon_{t-1}\varepsilon_{t-2} +$$

$$+ \theta_1\theta_2 E\varepsilon_{t-1}\varepsilon_{t-3} - \theta_2 E\varepsilon_{t-1}\varepsilon_{t-2} + \theta_1\theta_2 E\varepsilon_{t-2}^2 + \theta_2^2 E\varepsilon_{t-2}\varepsilon_{t-3} =$$

$$= -\theta_1(1-\theta_2)\sigma_\varepsilon^2$$

donde se ha utilizado en repetidas ocasiones la ortogonalidad íntertemporal del proceso de ruido blanco ε_t;. Del mismo modo, se tiene:

$$y_2 = E[y_t y_{t-2}] = E[-\theta_2\varepsilon_{t-2}\varepsilon_{t-2}] = -\theta_2\sigma_\varepsilon^2$$

mientras que y_k= 0 para todo k > 2. Por tanto, la FAS es:

$$\rho_1 = -\theta_1(1-\theta_2)/1+\theta_1^2+\theta_2^2 \text{ -}$$

$$\rho_2 = -\theta_2/1+\theta_1^2+\theta_2^2$$

$\rho_k = 0$ Para todo k>2

La identificación de un modelo MA(2) mediante su FAP es bastante difícil al igual que ocurre con las FAS de un AR(2), mientras que es muy sencillo identificar un modelo MA(2) por medio de su FAS igual que identificar un AR(2) mediante su FAP. Además los criterios de identificación son los mismos:

La FAS de un proceso MA(2) es cero para k > 2, de igual modo que la FAP

de un modelo AR(2) es cero para k> 2. Por otra parte, estos resultados son perfectamente generalizables para modelos de orden superior, AR (p) y MA (q), con p, q enteros positivos cualquiera.

2.1.5.MODELOS ARMA

Hasta ahora hemos visto modelo de series temporales sencillos, en los que un proceso estocástico tenía una estructura autorregresiva «pura», o una estructura de medias móviles «pura». Sin embargo, en el análisis empírico de series temporales económicas es muy frecuente encontrar representaciones que tienen una componente autorregresiva así como una componente de medias móviles. Estos modelos se denotan como modelos ARMA(p, q), donde p y q denotan, respectivamente, los órdenes de los componentes autorregresivo y de medias móviles.

Definición 2.12. La estructura ARMA más sencilla es el modelo ARMA(1, 1):

$$y_t = \delta + \phi y_{t-1} + \varepsilon_t - \theta \varepsilon_{t-1}$$

que depende de cuatro parámetros desconocidos: $\phi, \theta, \sigma_\varepsilon^2$ y δ. El proceso ARMA(1, 1) es estacionario cuando $|\phi| < 1$, e invertible cuando $|\theta| < 1$. Distintos modelos pertenecientes a esta familia pueden escribirse sin más que

variar los órdenes (p, q) de los componentes autorregresivo y de medias móviles.

La esperanza del proceso ARMA(1, 1) es $E(y_t) = \delta/(1-\phi)$. Para determinar su varianza, supongamos que δ = 0 lo que sin duda no afecta a la varianza del proceso, para obtener:

$$Var(\gamma_t) = \gamma_0 = E\left[(\phi y_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1})^2\right] = \phi^2 E(y_{t-1}^2) + 2\phi E(y_{t-1}\varepsilon_t) -$$

$$- 2\theta\phi E(y_{t-1}\varepsilon_{t-1}) + E(\varepsilon_t^2) - 2\theta E(\varepsilon_t\varepsilon_{t-1}) + \theta^2 E(\varepsilon_{t-1}^2) =$$

$$= \phi^2\gamma_0 - 2\phi\theta E(y_{t-1}\varepsilon_{t-1}) + \sigma_t^2 + \theta^2\sigma_t^2$$

donde se han utilizado las siguientes propiedades: a) Si $|\phi| < 1$, el proceso ARMA(1, 1) es estacionario y_{t-1} depende de ε_{t-1} anteriores, pero no de sus valores futuros y, en particular, y_{t-1} es independiente de ε_t y, b) por ser ruido blanco, $E(y_{t-1}\varepsilon_{t-1}) = 0$

Como además, $E(y_{t-1}\varepsilon_{t-1}) = \sigma_\varepsilon^2$, se tiene de la expresión anterior que

$$y_0(1-\phi^2) = \sigma_\varepsilon^2(1+\theta^2-2\phi\theta)$$

y, finalmente:

$$y_0 = [(1+\theta^2-2\phi\theta)/(1-\phi^2)]\sigma_\varepsilon^2$$

Los distintos valores de la función de autocovarianza del proceso ARMA(1, 1) pueden obtenerse de un modo similar:

$$\gamma_1 = E\left[y_{t-1}(\phi y_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1})\right] = \phi\gamma_0 - \theta\sigma_\varepsilon^2 = (1-\phi\theta)(\phi-\theta)/1-\phi^2 * \sigma_\varepsilon^2$$

$$\gamma_2 = E\left[(y_{t-2}(\phi y_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1})\right] = \phi\gamma_1$$

y, en general, $\gamma_k = \phi\gamma_{k-1}$ para todo k$\geq$ 2.

Como resultado, la función de autocorrelación de un proceso **ARMA (1, 1)** es

$$\rho_1 = (1-\phi\theta)(\phi-\theta)/1+\theta^2 - 2\phi\theta$$

$$\rho_k = \phi\rho_{k-1} \qquad \text{para todo } k \geq 2$$

En consecuencia, la función de autocorrelación de un proceso ARMA(1, 1) comienza del valor ρ_1 que acabamos de mostrar y, a partir de él, decrece a una tasa ϕ. Es decir, dicha función de- autocorrelación se comporta, a partir de k=1, como la función de autocorrelación de un proceso AR(1). Esta propiedad puede generalizarse: La función de autocorrelación de un proceso ARMA(p, q) se comporta como la función de autocorrelación del proceso AR(p), a partir de k > q.

Ello hace que la identificación de estos modelos por inspección de la serie temporal correspondiente no se ajuste a unas normas tan bien definidas como la identificación de modelos AR(p) o MA(q). Ello se debe a que tanto la función de autocorrelación como la función de autocorrelación parcial de los modelos ARMA heredan características de ambos componentes. Así puede

probarse también que mientras que la función de autocorrelación parcial de un modelo AR(1) es cero para k> 1, sin embargo, la de un modelo ARMA(1, 1) no tendrá esta característica, pues a ella hay que superponer la de la componente MA(1), que genera, como sabemos, una función de autocorrelación parcial que converge exponencialmente a cero. Las mismas consideraciones pueden hacerse para la función de autocorrelación simple.

Sin embargo, en la práctica esta dificultad no resulta excesivamente importante, pues tampoco se trata de obtener la mejor identificación del modelo en un primer intento. Así, el modo más frecuente de terminar con una especificación ARMA(2, 1), por ejemplo, es haber comenzado con una especificación AR(2), y observar que los residuos obtenidos tras la estimación de tal modelo tienen una estructura MA(1) (o recíprocamente).

Es interesante hacer hincapié en que esta superposición de modelos tiene un fundamento analítico. Si se ha especificado el modelo

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + u_t$$

y tras estimarlo e inspeccionar la función de autocorrelación de los residuos se observa que éstos parecen seguir una estructura MA(1):

$$\hat{u}_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

entonces estos dos modelos pueden unirse para obtener:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

es decir, una estructura ARMA(2, 1). En uno de los ejercicios al final del capítulo se pide al lector que lleve a cabo varios ejemplos similares a éste.

2.1.6. VARIABLES NO ESTACIONARIAS

Al presentar los modelos AR(p), supusimos que se satisfacían las condiciones necesarias para garantizar que el proceso en estudio era estacionario. Sin embargo, las series de datos económicos suelen caracterizarse por ser claramente no estacionarias. Como ya hemos citado, la observación de una tendencia lineal o cuadrática en el tiempo basta para rechazar el supuesto de estacionariedad.

Cuando esto ocurre en series económicas, es también cierto que si se toman primeras o segundas diferencias de la serie $\nabla y_t = y_t - y_{t-1}$ o $\nabla^2 y_t = y_t - 2y_{t-1} + y_{t-2}$ se obtienen series transformadas que son estacionarias.

Lo que se hace en tales situaciones es trabajar con las series en primeras o segundas diferencias, especificando y estimando un modelo para ellas. Como veremos más adelante, si se ha llevado a cabo un análisis de predicción para estas series, es bastante sencillo traducir las predicciones obtenidas para estas series diferenciadas en predicciones para las series

originales, que son, en definitiva, aquellas en cuyo análisis estaba interesado el investigador.

Los procesos no estacionarios que pueden transformarse en estacionarios mediante sus diferencias de orden *d* se conocen como *procesos estocásticos no estacionarios, homogéneos de orden d.* Un ejemplo de tales procesos es la caminata aleatoria. Su varianza crece con el tiempo, lo que le hace no estacionario. Sin embargo, su primera diferencia es, por definición:

$$\nabla y_t = y_t - y_{t-1} = \varepsilon_t$$

un ruido blanco y, como tal, es un proceso estocástico estacionario.

Cuando se especifica un modelo ARMA(p, q) para la variable $\Delta^d y_t$ se dice que tenemos un modelo ARIMA(p, d, q) para y_t.

Si la evidencia de no estacionariedad proviene de que el gráfico de la serie temporal que se pretende analizar muestra fluctuaciones cuya amplitud cambia para distintos intervalos del período muestral, pensaremos igualmente que el proceso estocástico que generó la serie temporal no es estacionario. En este caso, la no estacionariedad surgiría porque la varianza de las diferentes variables aleatorias que lo integran a lo largo del tiempo no son iguales entre si. El procedimiento habitual en esta situación consiste en transformar la variable tomando logaritmos, y pasar a analizar esta variable transformada. Veremos en la sección dedicada a predicción cómo recuperar

las predicciones de la serie original a partir de las predicciones obtenidas para la serie en logaritmos. Es importante hacer notar que la transformación logarítmica no va a corregir el problema de heteroscedasticidad sino que simplemente lo va a amortiguar, hasta el punto de hacerlo apenas perceptible. En este sentido, esta transformación es conceptualmente diferente de la diferenciación en el caso de no estacionariedad en la esperanza matemática que hemos visto antes. Por otra parte, la transformación logarítmica persigue estabilizar la varianza de la variable, mientras que la diferenciación busca estabilizar su esperanza matemática.

También debe notarse que la transformación Box-Cox incluye a la logarítmica como caso particular y que, en ocasiones, podría considerarse otra transformación de la familia Box-Cox.

2.1.7.ESTACIONARIEDAD E INVERTIBILIDAD -

Hay varias razones importantes para pretender que un proceso estocástico con el que se va a efectuar trabajo empírico sea estacionario: En primer lugar.

Si no es estacionario en media, por ejemplo, ello quiere decir que la esperanza matemática de las variables del proceso cambia con el tiempo, y

entonces habría que estimar un número infinito de parámetros. Aún más importante. Un proceso puede ser no estacionario porque sus momentos, o su distribución no existan. Lo mismo ocurriría con la varianza.

Un modelo AR(1) puede desarrollarse:

$$y_t = \phi y_{t-1} + \varepsilon_t = \phi^2 y_{t-2} + \phi\varepsilon_{t-1} + \varepsilon_t = \cdots =$$

$$= \phi^t y_0 + \phi^{t-1}\varepsilon_1 + \cdots + \phi\varepsilon_{t-1} + \varepsilon_t$$

Vimos anteriormente que el interés de que un proceso AR(1) sea estacionario se basa en que, de lo contrario, con $|\phi| > 1$ -se tendría que la expresión anterior es divergente, como puede apreciarse en el hecho de que los coeficientes de las variables aleatorias en la serie son crecientes sin límite.

Estos comentarios pueden extenderse a cualquier proceso AR(k). Por ejemplo, un proceso AR(2), $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + u_t$ es estacionario si al descomponer su ecuación característica en factores

$$1 - \phi_{1L} - \phi_2 L^2 = (1 - \lambda_1 L)(1 - \lambda_2 L)$$

donde L denota al operador de retardos, ambos parámetros, λ_1 y λ_2, tienen módulo menor que 1. Bastaría, por el contrario, que uno de ellos fuese mayor que 1 en valor absoluto, para que la realización de y_t, dependiese de

valores futuros de las variables ε_t, y quizá también de sus valores pasados. En la Sección 2.1.3 que las condiciones sobre los parámetros para que ello no ocurra son:

$$|\phi_2| < 1, \phi_2 + \phi_1 < 1, \phi_2 - \phi_1 < 1$$

Por otra parte, todo proceso MA de orden finito, por ser una combinación lineal de un número finito de variables aleatorias con distribución Normal, tiene asimismo una distribución Normal, con esperanza cero y una varianza que depende de los coeficientes del modelo, pero que es independiente del tiempo y es, por tanto, siempre un proceso estocástico estacionario.

En el caso de procesos de medias móviles, las condiciones similares a las de estacionariedad son las de invertibilidad, pero tienen una interpretación diferente: Cuando un proceso de medias móviles es invertible, entonces dicho modelo admite una representación autorregresiva, en que los valores pasados de la variable y_t, reciben una ponderación (coeficientes) cada vez menor cuanto más alejados en el tiempo. Esta propiedad de un modelo de medias móviles es fundamental a efectos de predicción, para lo que, como veremos, es necesario invertir primero el proceso MA y obtener su representación AR.

Un proceso ARMA (p, q) es estacionario si lo es su componente AR, y es

invertible si lo es su componente MA.

2.1.8.ESTACIONALIDAD

Las series temporales de datos económicos presentan generalmente características estacionales cuando se observan a una frecuencia inferior a la anual, ya sea mediante datos trimestrales o mensuales. Estas características se deben a que las decisiones tomadas por los agentes económicos en un determinado trimestre del año pueden estar correlacionadas con las decisiones tomadas en el mismo trimestre de otros años. Algo similar ocurre, en general, con datos mensuales.

Tales correlaciones pueden representarse analíticamente por modelos uni-variantes. Así, la especificación de las características estaciónales de una serie pudiera ser del tipo

$$y_t = \phi_{12} y_{t-12} + e_t$$

si además existen correlaciones entre las observaciones de meses consecutivos, se tendrá:

$$e_t = \phi e_{t-1} + \varepsilon_t$$

y, en resumen:

$$(1-\phi L)(1-\phi_{12}L^{12})y_t = (yt - \phi_{12}y_{t-12}) - \phi(y_{t-1} - \phi_{12} - y_{t-13)} = \varepsilon_t$$

La especificación del modelo univariante para la *estructura estocástica estacional* de una serie temporal se lleva a cabo para la identificación de un modelo para su *estructura estocástica regular*, con la salvedad de que para ello se examinan los valores estacionales de las funciones de autocorrelación, $\rho_4, \rho_8, \rho_{12..}$ si los datos son trimestrales, o $\rho_{12}, \rho_{24}, \rho_{36}$ **...** si son mensuales.

Así, por ejemplo, una serie puede requerir diferencias de orden estacional si los valores estacionales de su función de autocorrelación no tienden a cero rápidamente. Si trabajando con datos trimestrales se tiene que tanto ρ_4 como $\rho_8, \rho_{12..}$ son significativamente diferentes de cero, será en general apropiado transformar la variable mediante la diferencia:

$$z_t = \nabla_4 y_t = y_t - y_{t-4}$$

Así, una estructura posible para una serie mensual podría ser.-

$$\nabla\nabla_{12} \ln y_t = (1-\theta L)(1-\Theta L^{12})\varepsilon_t$$

que indica que la serie precisó de la transformación logarítmica, así como de una diferencia, tanto de carácter regular como estacional, esta última sugerida, sin duda, porque los valores de la FAS para k **=** 12, 24, 36 no convergían a cero. Tras estas transformaciones, los únicos valores no nulos de la FAS eran los correspondientes a k **=** 1 y k **=** 12, por lo que se especificó el modelo descrito.

2.1.9.PREDICCION CON MODELOS ARIMA

Es un resultado conocido en Estadística que cuando se pretende que la predicción $\hat{y}_{T+k}^{T}$ elaborada en el instante T acerca del valor de la variable endógena en el instante T+k minimice la función de pérdida

$$E_{T}\left[\left(y_{T+k}-\hat{y}_{T+k}^{T}\right)^{2}\right]$$

entonces la predicción debe ser $\hat{y}_{T+k}^{T}=E_{T}y_{T+k}$ es decir precisamente la esperanza condicional de la variable y_{T+k} calculada sobre la base de la información disponible en el instante T. Este es el criterio que utilizaremos en esta sección.

2.1.9.a. Modelos autorregresivos

En un modelo AR(1) se tiene:

$$E_{T}y_{T+1}=E_{T}\left(\delta+\phi y_{T}+\varepsilon_{T+1}\right)=\delta+\phi y_{T}+E_{T}\varepsilon_{T+1}=\delta+\phi y_{T}$$

donde hemos utilizado dos resultados:

a) Que y_{T} está en el conjunto de información disponible en el instante T. es decir, que la persona que realiza la predicción observa el valor

y_T previamente al cálculo de la predicción, con lo que $E_T y_T = y_T$

b) $E_T \varepsilon_{T+1} = 0$ por ser ε_t un ruido blanco.

De modo similar: $E_T y_{T+2} = E_T(\delta + \phi y_{T+1} + \varepsilon_{T+2}) = \delta + \phi E_T y_{T+1} + E_T \varepsilon_{T+2} =$

$$= \delta + \phi E_T y_{T+1} = \delta(1+\phi) + \phi^2 y_T$$

$$E_T y_{T+k} = \phi^k y_T + \delta(1 + \phi + \phi^2 + \cdots + \phi^{k-1})$$

Notemos que:

$$1 + \phi + \phi^2 + \cdots + \phi^{k-1} = (1 - \phi^k)/(1 - \phi)$$

y recordando que $Ey = \delta/(1-\phi)$ se tiene:

$$\delta(1 + \phi + \phi^2 + \cdots + \phi^k) = \left[(1 - \phi^k)/(1 - \phi)\right](1 - \phi)Ey = (1 - \phi^k)Ey$$

por lo que se tiene finalmente:

$$E_t y_{T+k} = \phi^k y_T + (1 - \phi^k)Ey$$

de modo que la predicción óptima k períodos hacia el futuro es una combinación lineal convexa de dos términos:

a) la última observación obtenida de la variable y_T ;

b) su esperanza matemática.

Según avanzamos hacia el futuro, la última observación recibe una

ponderación más pequeña, y la esperanza matemática recibe un peso más importante. Ello refleja el hecho de que, en un proceso estacionario, cuanto más lejos hacia el futuro queremos predecir, mayor será la incertidumbre bajo la que se obtiene la predicción: En un régimen de total incertidumbre, la predicción óptima del valor de una variable aleatoria es igual a su esperanza matemática. Por el contrario, cuando predecimos a horizontes cortos, la información muestral permite mejorar la predicción que se haría si se utilizase únicamente la esperanza matemática de la variable aleatoria. La propiedad característica del proceso AR(1) es que toda la información muestral relevante para la predicción quede resumida en e! último valor observado de la variable; en consecuencia, dicho valor es todo lo que se precisa, junto con la esperanza matemática del proceso, para elaborar predicciones.

En un proceso AR(2) se tiene:

$$E_T y_{T+1} = E_T(\delta + \phi_1 y_T + \phi_2 y_{T-1} + \varepsilon_{T+1}) = \delta + \phi_1 y_T + \phi_2 y_{T-1}$$

$$E_T y_{T+2} = \delta + \phi_1 E_T y_{T+1} + \phi_2 y_T = \delta(1+\phi_1) + (\phi_1^2 + \phi_2) y_T + \phi_1 \phi_2 y_{T-1}$$

$$E_T y_{T+3} = \delta(1+\phi_1+\phi_2+\phi_1^2) + \varphi_1(\phi_1^2 + 2\phi_2) y_T + (\phi^2{}_1 \phi^2{}_2 + \phi^2{}_2) y_{T-1}$$

y así sucesivamente. En procesos AR(p) de orden superior a 1 no existe una forma analítica sencilla para la predicción k períodos hacia el futuro. A pesar

de ello, es fácil observar una diferencia con respecto a la predicción con modelos AR(1), y es que la información muestral relevante para la predicción se resume en las p últimas observaciones de la variable a predecir.

2.1.9. b. Modelos de medias móviles

Un proceso MA(1) invertible puede escribirse:

$$y_{T+1} = -\theta y_T - \theta^2 y_{T-1} - \theta^3 y_{T-2} - ... + \varepsilon_{T+1}$$

por lo que:

$$E_T y_{T+1} = -\theta y_T - \theta^2 y_{T-1} - \theta^3 y_{T-2} + ...$$

$$E_T y_{T+2} = -\theta E_T y_{T+1} - \theta^2 y_T - \theta^3 y_{T-1} - \theta^4 y_{T-2} - ... =$$

$$= -(-\theta^2 y_T - \theta^3 y_{T-1} - \theta^4 y_{T-2} - ...) - \theta^2 y_T - \theta^3 y_{T-1} - \theta^4 y_{T-2} - ... = 0$$

$$E_T y_{T+k} = 0 \quad \text{para k > 2}$$

es decir, que la predicción de un proceso MA(1) para 2 o más períodos hacia el futuro es cero. Es fácil ver las razones que explican este resultado, puesto que el modelo:

$$y_t = \varepsilon_t - \theta \varepsilon_{t-1}$$

implica que

$$y_{T+1} = \varepsilon_{T+1} - \theta \varepsilon_T , \qquad y_{T+2} = \varepsilon_{T+2} - \theta \varepsilon_{T+1}$$

y así sucesivamente.

Como ε_t es ruido blanco, se tiene $E_T\varepsilon_{T+1}=0$ y más generalmente $E_T\varepsilon_{T+k}=0$, para todo k > 0. En consecuencia', $E_T y_{T+2}=0$ y también $E_T y_{T+s}=0$ para todo s>2.

Análogamente, puede verse que en un modelo MA(2) $E_T y_{T+k}=0$ para k >3. Más generalmente, en un modelo MA(q), $E_T y_{T+s}=0$ para todo s > q.

2.1.9. c. El modelo ARMA(1, 1)

Las expresiones analíticas correspondientes a las predicciones de valores futuros del modelo $y_t=\delta-\phi y_{t-1}+\varepsilon_t-\theta\varepsilon_{t-1}$ son:

$$E_{Tt}y_{T+1}=\delta+\theta y_T-\theta\varepsilon_T$$

$$E_T y_{T+2}=\delta+\theta E_T y_{T+1}=(1+\phi)\delta+\phi^2 y_T-\phi\theta\varepsilon_T$$

y, finalmente:

$$E_T y_{T+k}=(1+\phi+\phi^2+\ldots+\phi^{k-1})\delta+\phi^k y_T-\phi^{k-1}\theta\varepsilon_T$$

donde puede verse que, como ocurre para todo proceso estacionario:

$$\lim_{k\to\infty} E_T y_{T+k}=\delta/(1-\phi)=E(y_t)$$

la predicción converge a la esperanza matemática del proceso cuando el horizonte de predicción tiende a infinito.

2.1.9.d. Error de predicción

El error de predicción es la diferencia entre la realización de la variable aleatoria y la predicción hecha para dicho valor. Por supuesto, el error cometido en la predicción de y_{T+k} depende del período en que dicha predicción se hizo.

Denotamos por $e_T(k)$ el error de predicción k períodos hacia adelante, elaborada en el período T, es decir: $e_T(k) = y_{T+k} - E_T y_{T+k}$. En todos estos modelos, la esperanza del error de predicción es cero, precisamente, por ser el predictor óptimo la esperanza condicional de la variable a predecir. En efecto, se tiene:

$$E(e_T(k)) = E(y_{T+k} - E_T y_{T+k}) = -E[E_T y_{T+k}] = Ey_{T+k} - Ey_{T+k} = 0$$

Siendo el error de predicción $e_T(k)$ una variable aleatoria de esperanza cero, su varianza nos da una medida de su tamaño. En este sentido, una de las variables de más interés en el análisis de predicción es, por tanto, la varianza del error de predicción. Dicha varianza (más exactamente, su raíz cuadrada) es la cantidad que se utiliza para construir los intervalos de confianza de las predicciones puntuales obtenidas del modo descrito en los párrafos anteriores.

De acuerdo con dichas expresiones para las predicciones óptimas se tienen

los siguientes errores y varianzas de error de predicción (que el lector debe asegurarse de entender):

a) Errores de predicción en un modelo AR(1):

$$e_T(1) = y_{T+1} - E_T y_{T+1} = (\delta - \phi y_T + \varepsilon_{T+1}) - (\delta + \phi y_T) = \varepsilon_{T+1}$$

$$e_T(2) = \varepsilon_{T+2} + \phi \varepsilon_{T+1}$$

$$e_T(k) = \varepsilon_{T+k} + \phi \varepsilon_{T+k-1} + \phi^2 \varepsilon_{T+k-2} + ... + \phi^{k-1} \varepsilon_{T+1}$$

Varianza del error de predicción:

$$Var(e_T(1)) = \sigma_\varepsilon^2$$

$$\mathrm{Var}(\mathrm{e}_{\mathrm{T}}(2)) = \sigma_{\varepsilon}^{2}(1+\varphi^{2})$$

$$Var(e_T(3)) = \sigma_\varepsilon^2 (1 + \phi^2 + \phi^4)$$

$$Var(e_T(k)) = \sigma_\varepsilon^2 (1 + \phi^2 + \phi^4 + ... + \phi^{2(k-1)}) = \sigma^2{}_k [(1 - \phi^{2k})/(1 - \phi^2)$$

b) Errores de predicción en un modelo AR(2):

$$e_T(1) = \varepsilon_{T+1}$$

$$e_T(2) = \varepsilon_{T+2} + \phi_1 \varepsilon_{T+1}$$

$$e_T(3) = \varepsilon_{T+3} + \phi_1 \varepsilon_{T+2} + (\phi_1^2 + \phi_2) \varepsilon_{T+1}$$

Varianza del error de predicción

$$Var(e_T(1)) = \sigma_\varepsilon^2$$

$$Var(e_T(2)) = \sigma_\varepsilon^2(1+\phi_1^2)$$

$$Var(e_T(3)) = \sigma_\varepsilon^2(1+\phi_1^2+(\phi_1^2+\phi_2)^2)$$

c) Errores de predicción en un modelo MA(1):

$$e_T(1) = \varepsilon_{T+1}$$

$$e_T(2) = \varepsilon_{T+2} - \theta\varepsilon_{T+1}$$

$$e_T(k) = \varepsilon_{T+k} - \theta\varepsilon_{T+k-1}$$

para todo k>1

Varianza del error de predicción

$$Var(e_T(1)) = \sigma_\varepsilon^2$$

$$Var(e_T(2)) = \sigma_\varepsilon^2(1+\theta^2)$$

$$Var(e_T(k)) = \sigma_\varepsilon^2(1+\phi^2)$$

para todo k>1

e) Errores de predicción de un modelo ARMA(1,1)

$$e_T(1) = \varepsilon_{T+1}$$

$$e_T(2) = \varepsilon_{T+2} + (\phi-\theta)\varepsilon_{T+1}$$

$$e_T(3) = \varepsilon_{T+3} + (\phi-\theta)\varepsilon_{T+2} + \phi(\phi-\theta)\varepsilon_{T+1}$$

$$e_T(k) = \varepsilon_{T+k} + (\phi-\theta)\varepsilon_{T+k-1} + \phi(\phi-\theta)\varepsilon_{T+k-2} + \phi^2(\phi-\theta)\varepsilon_{T+k-3} + \ldots + \phi^{k-2}(\phi-\theta)\varepsilon_{T+1}$$

Varianza del error de predicción

$$Var(e_T(1)) = \sigma_\varepsilon^2$$

$$Var(e_T(2)) = \left[1 + (\phi - \theta)^2\right]\sigma_\varepsilon^2$$

$$Var(e_T(3)) = \left[1 + (\phi - \theta)^2 (1 + \phi^2)\right]\sigma_\varepsilon^2$$

$$Var(e_T(k)) = \left[1 + (\varphi - \theta)^2 [(1 - \phi^{2(k-1)}) / (1 - \phi^2)]\right]\sigma_\varepsilon^2$$

que, al igual que en los modelos AR(p) y MA(q), tiende á la varianza del proceso ARMA(1, 1) cuando el horizonte de predicción k tiende a infinito.

Obsérvese que las amplitudes del intervalo de confianza para las sucesivas predicciones elaboradas con un modelo AR crecen continuamente con el horizonte de predicción, mientras que para un modelo MA(q) permanecen constantes a partir de q períodos hacia el futuro. El lector puede notar asimismo que el error de predicción un período hacia el futuro es el mismo ε_{T+1}, independientemente de cual sea el verdadero modelo de la serie temporal en estudio. Este es un resultado lógico, pues dicho error es la componente del valor futuro de la serie que no puede predecirse sobre la base de la información muestral, lo que en la literatura de series temporales se conoce como innovación en la serie.

Ello no quiere decir, sin embargo, que el error de predicción de una serie

temporal un período hacia el futuro vaya a ser independiente del modelo utilizado para obtener dicha predicción. El error de predicción, por definición, es la diferencia entre la realización de la serie y su predicción óptima, obtenida con el mejor modelo que pueda elaborarse sobre la base de la información muestral. En la práctica, estas diferencias se reflejan en diferentes valores estimados del parámetro σ_{ε}^{2} según se ajuste un modelo u otro. Dicho de otro modo, si no se utiliza el mejor modelo para la serie, entonces el error de predicción incluirá, junto con la innovación ε_t, un error de especificación ξ_t y, agregadamente, tendrán una varianza superior a la de la auténtica innovación.

2.1.9.e. Intervalos de confianza para las predicciones

La varianza del error de predicción puede utilizarse para obtener intervalos de confianza de las predicciones elaboradas, mediante la expresión:

$$E_T y_{T+k} \pm \lambda_a \hat{\sigma}_{eT(k)}$$

donde, si se supone que la innovación ε_t, sigue una distribución Normal, el parámetro λ_a se obtendrá de las tablas de dicha distribución, al nivel de

confianza α elegido. Así, por ejemplo, para obtener en el instante T un intervalo del 95 por 100 de confianza para el valor del proceso y_T, en el instante T + k, supuesto que la innovación ε_{T+k} siga una distribución Normal, basta sumar y restar 1,96 veces el valor estimado de $\sigma_{eT(k)}$ a la predicción $E_T y_{T+k}$ Como es habitual, este intervalo será tan sólo una aproximación, por haber estimado el parámetro $\hat{\sigma}_e$.

2.1.9.f. Predicción de una serie en diferencias

Si se ha estimado un modelo ARIMA con un número no nulo de diferencias, entonces será preciso recuperar las predicciones de la serie original a partir de las predicciones elaboradas para la serie en diferencias. Ello puede hacerse del siguiente modo: Supongamos que y_T , denota la serie en cuyo análisis estamos interesados, y que se ha especificado y estimado un modelo univariante para la serie de primeras diferencias: $z_t = \nabla y_t$. Entonces, es claro que:

$$E_T z_{T+k} = E_T y_{T+k} - E_T y_{T+k-1}$$

por lo que:

$$E_T y_{T+k} = E_T z_{T+k} - E_T y_{T+k-1} = \ldots =$$

$$= E_T z_{T+k} + E_T z_{T+k-1} + E_T z_{T+k-2} + ... + E_T z_{T+1} + y_T$$

y, consecuentemente, para obtener las predicciones de la serie y_t, basta añadir a su última observación muestral las predicciones elaboradas para sus incrementos, z_{T+k}

El lector puede comprobar, como se pide en uno de los ejercicios al final del capítulo, que la recuperación de las predicciones de una serie y_t, a partir de las predicciones elaboradas para la serie de segundas diferencias $\nabla^2 y_t$ puede llevarse a cabo de un modo similar al que acabamos de describir.

Por último cabe citar que si el modelo univariante se ha estimado para la transformación logarítmica de la variable original $Z_t = \ln(y_t)$ entonces el modo de recuperar las predicciones de los valores futuros de y_t es:

$$E_T Y_{T+k} = \exp\left\{ E_T z_{T+k} + \frac{1}{2} Var(e_T(k)) \right\}$$

Dependiendo del tamaño de la varianza del error de predicción, la expresión anterior supondrá una diferencia significativa o no con respecto a la simple alternativa de hallar la función exponencial de los valores $E_T z_{T+k}$

Sin embargo, los límites inferior y superior de los intervalos Qe confianza

para $E_T y_{T+k}$ deben hallarse calculando los valores de la función exponencial en los limites inferior y superior del intervalo de $E_T z_{T+k}$. En consecuencia, el intervalo de confianza que se obtenga para la predicción $E_T y_{T+k}$ no será simétrico alrededor de dicha predicción.

Nótese, por otra parte, que las variaciones en las predicciones de z_{T+k} para valores sucesivos de k pueden interpretarse como variaciones porcentuales previstas en la variable original y_{t+k}

2.1.10. ESTIMACION DE MODELOS ARIMA

2.1.1O.a. Estimación de modelos autorregresivos

Un modelo autorregresivo presenta una diferencia notable con respecto a los modelos econométricos que hasta ahora hemos considerado. Las variables explicativas son ahora aleatorias, ya que son retardos de la variable y_t que es aleatoria. Puede probarse que, en tales condiciones, el estimador MCO tiene buenas propiedades y, en particular, es un estimador consistente

siempre que las variables explicativas x_{it}, satisfagan la condición:

$$E(x_{i,t-s}\, u_t) = 0$$

Si el término de error no tiene autocorrelación y si e! modelo es estacionario, esta propiedad se satisface. En efecto, los valores $x_{i,t-s}$ de la condición anterior son ahora retardos de y_t t, mientras que u_t, es ahora ε_t. Bajo el supuesto de estacionariedad, la variable y_t, depende de ε_t , y de sus valores anteriores, pero de ningún valor futuro de ε_t. Por tanto, las esperanzas $E(y_{t-s}\varepsilon_t)$ son cero para todo s> 0.

En tal caso, la estimación consistente del modelo autorregresivo

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t$$

puede llevarse a cabo por MCO. Por tanto, todo depende de que la especificación sea correcta.

Si, por el contrario, el término de error del modelo tuviese autocorrelación, entonces la condición de ortogonalidad no se cumpliría y el estimador MCO dejaría de ser apropiado. En efecto, si ε_t, no fuese ruido blanco sino que obedeciese, por ejemplo, al modelo $\varepsilon_t = \phi\varepsilon_{t-1} + \xi_t$ con ξ_t ruido blanco, entonces se tendría que: a) y_{t-1} estaría correlacionado con ε_{t-1}, a través del modelo univariante elaborado, y b) ε_t también estaría correlacionado con, ε_{t-1} a través de su modelo de autocorrelación. En estas

condiciones $E(y_{t-1}\varepsilon_t) \neq 0$, ya que ambos están correlacionados con ε_{t-1} contradiciendo la condición de ortogonalidad necesaria para justificar el uso del estimador MCO.

La autocorrelación del término de error de un modelo univariante es un indicio evidente de mala especificación de dicho modelo. Una especificación correcta debe generar un término de error con estructura de ruido blanco.

2.1.10.b. Estimación de modelos de medias móviles

Como ejemplo de la estimación de modelos de medias móviles, discutiremos las cuestiones importantes en el contexto de un modelo MA(2): $y_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2}$**.** Para ello, si suponemos que la innovación ε_t sigue una distribución Normal $(0, \sigma_\varepsilon^2)$, se tiene la siguiente expresión para el logaritmo de la función de verosimilitud:

$$LogL(\theta_1, \theta_2, \sigma_\varepsilon^2) = Cons\tan te - (t)/(2)\log\sigma_\varepsilon^2 - \left[SR(\theta_1, \theta_2)\right]/(2\sigma_\varepsilon^2)$$

$$LogL(\theta_1, \theta_2, \sigma_\varepsilon^2) = \text{Constante} - \frac{t}{2}\log\sigma_\varepsilon^2 - \frac{\left[SR(\theta_1, \theta_2)\right]}{2\sigma_\varepsilon^2}$$

donde $SR(\theta_1, \theta_2)$ denota la suma residual: $SR(\theta_1, \theta_2) = \sum_3^T \varepsilon_t^2$ de modo que, condicional en el valor del parámetro **,** σ_ε^2 la maximización de la función de

verosimilitud coincide con la minimización de la suma residual. Se trataría, por tanto, de minimizar la suma de cuadrados de los residuos:

$$\text{MIN} \quad SR(\theta_1, \theta_2) = \sum_3^T \varepsilon_t^2 = \sum_3^T (y_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2})^2$$

donde, como puede verse, se están ignorando los dos primeros períodos de la muestra, pues no se dispone para ellos de observaciones acerca del ruido blanco ε_t. La solución a este problema de minimización proporciona los valores estimados de los parámetros del modelo y, finalmente, la varianza σ_ε^2 se estima mediante:

$$\sigma_\varepsilon^2 = \frac{SR(\hat{\theta}_1, \hat{\theta}_2)}{T - p - q}$$

donde p + q es el número de parámetros estimados en el modelo univariante. en este caso 2.

La minimización de la suma residual anterior podría llevarse a cabo mediante una red de búsqueda. Este procedimiento sería muy sencillo en el caso de un proceso MA(1), puesto que el espacio paramétrico admisible sería (—1., 1), y bastaría hacer una partición de este intervalo, generar la serie de residuos $\{\hat{\varepsilon}_t\}$ tomando como dadas las observaciones Y_1 e y_2 y evaluando la función SR(θ). La certeza de haber hallado un mínimo global en vez de un mínimo local, aumenta si la partición que se lleva a cabo es

suficientemente fina, o si se lleva a cabo una exploración bastante exhaustiva del espacio paramétrico. En el caso de un proceso MA(2), el proceso de búsqueda es más complejo, pues el espacio paramétrico tiene dos dimensiones. Sin embargo. dicho espacio queda limitado por las condiciones de invertibiidad del proceso que vimos anteriormente.

Como alternativa, dados unos valores iniciales de los parámetros la expresión $\varepsilon_t = y_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}$ podrá utilizarse para generar observaciones de ε_t, para t= 3, 4, ..., T.

Claramente, los valores numéricos de los residuos $\hat{\varepsilon}_t \geq 3$ varían con los valores de los parámetros $\theta_1 y \theta_2$ y se trata de encontrar el par de valores θ_1, θ_2

para el que la suma de cuadrados de los residuos es mínima. Es imposible utilizar todos los pares de valores paramétricos posibles y calcular para cada uno de ellos la suma residual correspondiente, por lo que es preciso recurrir a un procedimiento analítico, que pasamos a describir.

Denotamos por $\theta_1^0 \theta_2^0$ los valores iniciales de los parámetros y por ε_t^0, t>3 los residuos obtenidos con el par de valores $\theta_1^0 \theta_2^0$. EL verdadero valor del residuo ε_t; es, sin embargo, el correspondiente a los verdaderos valores de los parámetros $\theta_1 y \theta_2$ y se puede llevar a cabo la siguiente aproximación por

medio del desarrollo en serie de Taylor:

$$\varepsilon_t = \varepsilon_t^0 + (\theta_1 - \theta_1^0)\left[\partial\varepsilon_t / \partial\theta_1\right]_{\theta=\theta^\circ} + (\theta_2 - \theta_2^0)\left[\partial\varepsilon_t / \partial\theta_2\right]_{\theta=\theta^\circ} +$$

$$+1/2(\theta_1 - \theta_1^0)^2\left[\partial^2\varepsilon_t / \partial\theta^2{}_1\right]_{\theta=\theta^\circ} + 1/2(\theta_1 - \theta_1^0)\left[\partial^2\varepsilon_t / \partial\theta_1\partial\theta_2\right]_{\theta=\theta^\circ} +$$

$$+1/2(\theta_2 - \theta_2^0)^2\left[\partial^2\varepsilon_t / \partial\theta_2^2\right]_{\theta=\theta^\circ} + \ldots \qquad [2.1.7],$$

donde ya hemos despreciado los términos de orden superior a 2, en la confianza de que tanto las potencias $(\theta_1 - \theta_1^0)^3$ como las derivadas parciales de esos órdenes serán pequeñas en valor absoluto. Si por idénticas razones despreciamos asimismo los términos de segundo orden que aparecen en [2.1.7], se llega a

$$\varepsilon_t = \varepsilon_t^0 + (\theta_1 - \theta_1^0)\left[\partial\varepsilon_t / \partial\theta_1\right]_{\theta=\theta^\circ} + (\theta_2 - \theta_2^0)\left[\partial\varepsilon_t / \partial\theta_2\right]_{\theta=\theta_2^0}$$

y puesto que en este modelo

$$\partial\varepsilon_t / \partial\theta_1 = \varepsilon_{t-1} \qquad \partial\varepsilon_t / \partial\theta_2 = \varepsilon_{t-2}$$

se tiene:

$$\varepsilon_t^0 - \theta_1^0\varepsilon_{t-1}^0 - \theta_2^0\varepsilon_{t-2}^0 = -\theta_1\varepsilon_{t-1}^0 - \theta_2\varepsilon_{t-2}^0 + \varepsilon_t$$

que muestra que se pueden obtener estimaciones de los parámetros $\theta_1 y \theta_2$ sin más que estimar el modelo lineal:

$$w_t = \theta_1 x_{1t} + \theta_2 x_{2t} + \varepsilon_t$$

Donde

$$w_t = \varepsilon_t^0 - \theta_1^0 \varepsilon_{t-1}^0 - \theta_2^0 \varepsilon_{t-2}^0, x_1 = -\varepsilon_{t-1}^0, x_{2t} = -\varepsilon_{t-2}^0 \qquad [2.1.8]$$

El lector debe notar que este procedimiento no es sino la aplicación del algoritmo Gauss-Newton a la minimización de la suma residual condicional.

La estimación del modelo [2.1.8] requiere de la previa construcción de estas variables. Los valores θ_i^0, i = 1, 2 que en ellas aparecen son los valores iniciales de- los parámetros que el investigador debe seleccionar. Pongamos, por ejemplo, que θ_i^0 = 0,10, aunque más adelante daremos unas reglas más rigurosas para la elección de estos valores iniciales. Por otra parte, la relación [2.1.6] puede utilizarse para generar «datos» para la variable ε_t^0 para **t=**3, 4,….,T, utilizando los valores iniciales de los parámetros. La estimación del modelo [2.1.8] se llevaría a cabo con T — 3 observaciones.

El procedimiento descrito puede y debe iterarse, utilizando como valores $\theta_1^0 \theta_2^0$ en una segunda etapa las estimaciones obtenidas en la primera etapa. El esquema iterativo podría terminar cuando las diferencias entre los valores paramétricos iniciales y finales de una etapa del proceso son pequeñas. En tal caso, se dice que el proceso de estimación numérica que hemos descrito ha convergido (supuestamente a los verdaderos valores de los parámetros).

Como posible criterio de convergencia, podría decidirse detener el proceso si las variaciones en todos los parámetros son inferiores a 0,001, o al 5 por 100 de su valor inicial.

Un criterio de convergencia alternativo consistiría en detener el proceso si la variación producida en la suma residual (recordemos que ésta es la función objetivo de! problema de optimización) es pequeña. De nuevo, habría que definir lo que se entiende por «pequeña», pero podría ser una diferencia inferior al 1 por 100. Un criterio más exigente sería el de aceptar la convergencia del proceso sólo cuando se cumplen simultáneamente los dos criterios que acabamos de citar.

Por supuesto, también existe la posibilidad de que el proceso anterior no converja. Ello puede deberse al hecho de que el procedimiento de estimación se basa en una aproximación lineal a la función que hace depender ε_t , de los parámetros del modelo, y no en la verdadera relación entre ambos. Otra posibilidad es que los valores iniciales escogidos para los parámetros no hayan sido muy adecuados por lo que, cuando esto ocurre, debe repetirse el procedimiento con otros valores iniciales. Finalmente, la convergencia puede no producirse por una mala especificación del modelo, es decir, porque el modelo univariante que se está tratando de estimar no sea el modelo lineal que mejor representa la estructura del proceso

estocástico que generó la serie temporal que está siendo objeto de análisis.

El procedimiento que acabamos de describir es válido, con las modificaciones obvias, para la estimación de cualquier modelo univariante. Al estimar el modelo

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - ... - \theta_q \varepsilon_{t-q}$$

se ignoran los primeros q residuos, lo que equivale a suponerlos iguales a su esperanza, cero. Por ello, este procedimiento de estimación se conoce como de mínimos cuadrados condicionados al supuesto de que los primeros residuos son iguales a cero. Si la muestra consta de pocas observaciones, el número

de grados de libertad (número de observaciones menos número de parámetros a estimar) puede ser muy pequeño, e ignorar las primeras observaciones en la estimación será una mala aproximación. -

2.1.1O.c. Obtención de valores iniciales para los parámetros del modelo

Para llevar a cabo el procedimiento analítico de estimación que hemos descrito en los párrafos anteriores es preciso haber identificado previamente

un modelo para la serie temporal en estudio. Esta identificación se habrá llevado a cabo mediante el análisis de sus funciones de autocorrelación y de autocorrelación parcial. Los valores numéricos de las estimaciones muestrales de estas funciones pueden utilizarse para obtener estimaciones iniciales de los parámetros del modelo.

Por ejemplo, supongamos que el examen de la función de autocorrelación parcial de una serie sugiere que el proceso que la ha generado tiene una estructura autorregresiva pura. En tal caso, las ecuaciones de Yule-Walker permiten obtener valores de los parámetros ϕ_i a partir de los valores estimados de la función de autocorrelación. Nótese que si las estimaciones muestra-les de la función de autocorrelación fuesen muy próximas a los verdaderos valores de dicha función, entonces la solución al sistema de Yule-Walker proporcionaría los valores de los parámetros del modelo, sin necesidad de utilizar el proceso analítico antes descrito. Sin embargo, la estimación de la función de autocorrelación no suele tener tan buenas propiedades, y las primeras iteraciones del procedimiento anterior producen, en general, variaciones significativas.

Si, por el contrario, se ha especificado un modelo MA(1), entonces, recordando que $p_1 = -(\theta)/(1+\theta^2)$ esta ecuación puede invertirse para obtener el valor de θ . La ecuación que así resulta es de segundo grado y tiene, por

tanto, dos soluciones; en general, una de las raíces es mayor, en valor absoluto, y la otra menor que 1. Para garantizar la invertibilidad del proceso, escogemos siempre la raíz inferior a la unidad en valor absoluto.

Para procesos de orden superior, la obtención de valores iniciales de los parámetros es más compleja, pero puede intentarse un procedimiento similar al descrito.

Una alternativa consistiría en obtener la representación autorregresiva del proceso MA, y estimarla. Este método tiene la gran ventaja de su sencillez. Sin embargo, ya hemos visto que la representación AR de cualquier proceso MA de orden finito es de orden infinito. Por tanto, habría que truncar la autorregresión. Ello no es un problema excesivo, puesto que los coeficientes en los sucesivos retardos son potencias de los coeficientes ϕ**,** que son menores que 1 en valores absoluto. Pero además, una vez truncada la regresión, los coeficientes restantes están sujetos a restricciones no lineales, que habría que imponer si se pretende obtener un estimador eficiente. Para ilustrar esta problemática, consideremos el modelo MA(1), cuya representación autorregresiva es:

$$y_t = \varepsilon_t - \theta y_{t-1} - \theta^2 y_{t-2} - \theta^3 y_{t-3} - \cdots$$

Si el parámetro θ fuese suficientemente pequeño como para que

$$y_t = -\theta y_{t-1} + u_t$$

fuese una buena aproximación, entonces el coeficiente de y_{t-1} sería una estimación bastante aproximada del parámetro θ.. Hay que hacer notar, sin embargo, que tal regresión no dejaría de ser una mala especificación de un modelo que es, realmente, una autorregresión de orden infinito. De este modo. el término de error u_t , incorporaría los retardos omitidos, y se produce el problema de correlación entre el regresor y_{t-1} y u_t , al que antes hicimos mención. También es cierto que la trascendencia de tal correlación depende de la magnitud de los coeficientes de los retardos omitidos, es decir, de la calidad de la aproximación anterior.

Dicha aproximación puede mejorarse, y con ello aminorar el sesgo producido por la correlación entre y_{t-1} y u_t, si se incluye algún otro retardo, por ejemplo:

$$y_t = -\theta y_{t-1} - \theta^2 y_{t-2} + u_t$$

pero como vemos, ya aparece una restricción no lineal entre los coeficientes de y_{t-1} e y_{t-2}

2.1.11. DIAGNOSTICO DEL MODELO

Tras estimar un modelo ARIMA, es esencial llevar a cabo un análisis de los

coeficientes y residuos del modelo, con el objetivo de detectar posibles indicios de mala especificación. Respecto a los coeficientes, éstos deben de satisfacer siempre las condiciones de estacionariedad e invertibilidad. Cuando se estiman términos AR(2) o MA(2), deben obtenerse siempre sus raíces, y factorizar dichos términos cuando las raíces sean reales, por la razón que en seguida veremos. También debe hacerse hincapié en que los coeficientes estimados serán relevantes si el algoritmo numérico de estimación ha convergido. Si no lo ha hecho, puede concederse un número mayor de iteraciones, pero puede también ser un indicio de ausencia de estacionariedad o de invertibilidad, por tanto, de -un deficiente ajuste del modelo especificado a las series analizadas.

Asimismo, deben examinarse cuidadosamente los residuos resultantes hasta que se consiga, en la medida de lo posible, eliminar toda duda acerca de que éstos obedecen un proceso de ruido blanco. Esta verificación es crucial, pues sobre tal supuesto se habrá diseñado la estrategia de estimación y predicción. Como hemos expuesto en las secciones anteriores. Cualquier evidencia en contra de la hipótesis de ruido blanco-para -los-residuos constituye un indicio de mala especificación del modelo.

2.1.11.a. Análisis de residuos

Ya vimos en la Sección 2.1.5 cómo unos residuos MA(1) obtenidos en un modelo AR(2) sugieren un modelo más correcto ARMA(2, 1). Como otro ejemplo, supongamos que se ha especificado y estimado el modelo:

$$y_t = \delta + \phi_1 y_{t-1} + u_t \qquad [2.1.9]$$

cuyos residuos parecen obedecer el modelo: -

$$u_t = \phi_2 u_{t-1} + \varepsilon_t \qquad [2.1.10]$$

donde, a diferencia de u_t, ε_t, es ruido blanco. Una simple sustitución del modelo [2.1.10] en [2.1.9] muestra que el verdadero modelo de y_t, es:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \varepsilon_t$$

Donde $\beta_0 = \delta(1-\phi_2), \beta_1 = \phi_1 + \phi_2$ y $\beta_2 = -\phi_1\phi_2$ es decir, el proceso y_t, tiene una estructura AR(2) aunque, por error, se había especificado un modelo AR(1).

Sin embargo, si el término de error del modelo [2.1.9] tuviese una estructura de media móvil: $u_t = \varepsilon_t + \theta\varepsilon_{t-1}$ entonces la sustitución directa de esta estructura en [2.1.9] muestra que, en realidad, y_t, sigue una estructura ARMA(1, 1)

Supuesto que se ha identificado correctamente una estructura AR, la elección del orden se basa en contrastes de significación llevados a cabo con los valores estimados de la función de autocorrelación parcial. Cuando el verdadero modelo es AR(p), la distribución de los valores estimados de la FAP es, aproximadamente: $\hat{\phi}_{jj} \sim N(0, \frac{1}{T})$ para j**>** p, por lo que si $\hat{\phi}_{jj}$ está en el Intervalo $\pm 2/\sqrt{T}$ para todos los valores de j superiores a un cierto p, entonces no se rechazará la hipótesis de que el orden del modelo AR es menor o igual que p.

Un procedimiento similar permite identificar el orden de un proceso MA(q). Como sabemos, un proceso MA(q) se caracteriza porque su función de autocorrclación simple teórica es cero para valores k> q. Se trata, por tanto, de efectuar contrastes de hipótesis de significación estadística acerca de los valores estimados de dicha función. En un proceso de ruido blanco, las estimaciones r_k se distribuyen asintóticamente $N\left(0, \frac{1}{T}\right)$**.** Así, se acepta la no significación de las estimaciones r_k si caen dentro del intervalo $\pm 2/\sqrt{T}$. Si ello ocurre para todo k mayor que un cierto q, pensamos que el proceso es MA(q). Si ocurre para todo k, entonces se mantiene el supuesto de que la serie temporal de datos proviene de un proceso de ruido blanco. Si el proceso no es ruido blanco, entonces el cociente $1/\sqrt{T}$ puede no ser una

aproximación suficientemente buena a la desviación típica de r_k. Bartlett [1946] ha probado que si p_k es distinto de cero para k <q e igual a cero para k > q, entonces la varianza de r_k es aproximadamente igual a:

$$Var(r_k) = (1 + 2(p_1^2 + p_2^2 + ... + p_q^2))/T \text{ -}$$

En particular, para contrastar la hipótesis nula de que el proceso estocástico subyacente tiene estructura de ruido blanco, como ocurre cuando aplicamos este procedimiento a los residuos de un modelo estimado, la expresión anterior se reduce a:

$$Var(r_k) = 1/T$$

que es la justificación para utilizar $\pm 2/\sqrt{T}$ como intervalo de confianza del 95 por 100 para los valores estimados de la función de autocorrelación. Como acabamos de decir, tal práctica sólo está justificada si la hipótesis nula que se contrasta es la de ruido blanco. Si, por el contrario, ya se ha aceptado previamente que la serie en análisis proviene de un proceso con una cierta estructura estocástica, y se está considerando la posibilidad de añadir más estructura. Entonces la práctica anterior sólo puede entenderse como una aproximación sencilla al contraste dado por la expresión anterior para la $Var(r_k)$.

La fórmula de Bartlett es únicamente una aproximación a la verdadera

varianza de los valores estimados de la función de autocorrelación. Dicha aproximación tiende a sobrestimar el valor de la varianza, generando así unos intervalos de confianza de un tamaño mayor del que realmente debieran tener. En consecuencia, la utilización de la fórmula de Bartlett tiende a mantener la hipótesis nula de no significación de r_k más a menudo de lo que debiera y, con ello, a no detectar estructura estocástica en casos en que dicha estructura existe.

Conviene, por tanto, prestar especial atención a la posible detección de alguna regularidad en la función de autocorrelación que la asemeje a alguna de las vistas en este capítulo, incluso si sus valores no son significativos. En especial, conviene ser exigente con las estimaciones de los primeros valores de dicha función, y utilizar para ellos un intervalo de confianza de aproximadamente ± 1,25 o ± 1,5 desviaciones típicas, en vez de las habituales dos desviaciones típicas. En definitiva, $1/T$ no es sino una cota superior para la varianza de dichos valores.

En ocasiones, para formalizar el contraste de la hipótesis nula H_0: «los residuos del modelo son ruido blanco», se resume toda su estructura de auto-correlación en el estadístico de Box-Pierce:

$$Q = T(r_1^2 + r_2^2 + ... + r_k^2)$$

que se distribuye como una chi-cuadrado con k **-** p - a grados de libertad,

donde p y q son el número de parámetros AR y MA estimados. Con el objeto de ganar potencia, se ha propuesto también el estadístico modificado por Ljung y Box:

$$Q`= T(T+2)\sum_{1}^{k}(r_j^2)/(T-j)$$

que tiene distribución chi-cuadrado con k grados de libertad.

De igual modo, el gráfico de residuos puede ilustrar la presencia de posibles tendencias que no apareciesen obvias en la FAS. Recordemos que, generalmente, en situaciones de no estacionariedad, la FAS decrece hacia cero sólo lentamente, pero, en ocasiones, los residuos muestran una clara tendencia y, sin embargo, sólo los valores iniciales de la FAS están fuera del intervalo de confianza del 95 por 100.

Cuando se analiza una variable estacional, debe prestarse especial atención a los primeros valores de las FAS y FAP, pero también a tos primeros valores

de orden estacional. Así, con una serie trimestral, el analista debe observar detenidamente los valores 1, 2, 3, 4 y 8 de ambas funciones, a los valores 1, 2, 3, 4, 5, 6, 12 y 24 si trabaja con una serie mensual. Todo ello sin menoscabo de examinar asimismo el resto de los valores.

2.1.11 .b. Sobreparametrización y sobrediferenciación

Dos aspectos de la identificación de un modelo univariañte están muy relacionados entre si: a) la posible sobreparametrización, por existencia de factores comunes, y b) la posible sobrediferenciación del proceso. En primer lugar, debe notarse que los modelos:

$$\nabla^2 y_t = (1 - 0{,}97L)\varepsilon_t \qquad [2.1.11]$$

$$\nabla y_t = \varepsilon_t \qquad [2.1.12]$$

son prácticamente indistinguibles, puesto que como $\nabla = (1 - 1{,}0L)$ existe un factor común en ambos miembros de [2.1.11]. De modo similar, son modelos aproximados:

$$(1 - 0{,}95L)y_t = (1 - 0{,}20L)\varepsilon_t \qquad [2.1.13]$$

$$\nabla y_t = (1 - 0{,}20L)\varepsilon_t \qquad [2.1.14]$$

puesto que un término AR(1) con coeficiente positivo y próximo a 1 sugiere la no estacionariedad de la variable y, por ello, la conveniencia de diferenciarla. Por último, los modelos

$$(1-1{,}20L+0{,}35L^2)\nabla y_t=(1-0{,}70L)\varepsilon_t \qquad [2.1.15]$$

$$(1-0{,}50L)\nabla y_t=\varepsilon_t \qquad [2.1.16]$$

son indistinguibles, por cuanto que la factorización del polinomio AR(2): $(1-1{,}20L+0{,}35L^2)=(1-0{,}70L)(1-0{,}50L)$ muestra la existencia de un factor común en ambos miembros de [13.5] que no era evidente a simple vista. Como norma general, el investigador siempre debe obtener las raíces del polinomio cuando estima un término AR(2).

En ocasiones, no resulta evidente cuál es el orden correcto de diferenciación de una serie. A este respecto, no hay una doctrina clara: algunos investigadores prefieren no diferenciar por miedo a eliminar información relevante; otros temen que no diferenciar deje en la serie aspectos de no estacionariedad que sesguen el proceso de elaboración del modelo. Esta última estrategia parece, sin embargo, dar mucho mejores resultados. El analista no debe temer introducir una diferencia adicional en la variable, especialmente si incorpora simultáneamente un término de media móvil. Por ejemplo, pasaría del modelo:

$$(1+\phi_1 L+\phi_2 L^2)\nabla y_t=\varepsilon_t \qquad [2.1.17]$$

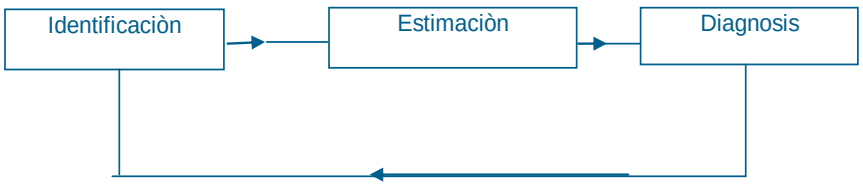
Sobre cuyos residuos existen sospechas de no estacionariedad, al modelo:

$$(1+\phi_1 L+\phi_1 L^2)\nabla y_t=(1-\theta L)\varepsilon_t \qquad [2.1.18]$$

y si el parámetro θ se aproxima mucho a 1, puede volverse al modelo [2.1.7]. Sin embargo, siempre que el valor de θ no produzca un modelo claramente no invertible, conviene mantener la especificación [2.1.8], pues la combinación de diferencia y término de media móvil suele producir buenos resultados predictivos.

Cuando se han detectado factores comunes, no debe simplemente mantenerse el modelo tras la simplificación pertinente, sino que debe volverse a estimar, al igual que ocurre con cualquier evidencia de mala especificación que se obtuviese de los residuos o cuando se decide introducir una diferencia adicional. En definitiva, el proceso de elaboración de un modelo univariante tiene varias etapas:

Gráfico 2.1.6.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Modelo Univariante



Autor: José Aguayo E. Fuente: Econometría 2da. Edición 2000 por A. Novales, Mc Grawhill

que sólo finalizan cuando la diagnosis de un modelo estimado no sugiere

indicación de mala especificación.

2.1.11. c. Valores influyentes y anomalías. Análisis de intervención

Otro aspecto importante en la elaboración de modelos univariantes lo constituye la posible existencia de valores influyentes. Estos son valores de la serie que se analiza que, por su magnitud, distorsionan el proceso iterativo de construcción del modelo. Piénsese que, en general, el modelo se especifica para la variable diferenciada al menos una vez, puesto que prácticamente ninguna variable stock, y tan sólo pocas variables flujo o precios, son estacionarias. Por consiguiente, un aumento o disminución mensual, trimestral o anual, según la frecuencia de observación de la serie, de importante magnitud, pueden resultar valores influyentes.

Hay dos posibles estrategias frente a este tipo de valores: Por una parte, pueden tratarse de igual modo que el resto de las observaciones de la serie. Por otro lado, puede estimarse su influencia para, en cierto modo, descontar su posible efecto en la elaboración del modelo univariante.

En algunos casos, una intervención de política económica tiene un importante efecto sobre la variación en la oferta monetaria o los tipos de interés

que podría hacer dichos valores influyentes. Precisamente porque se sabe que han sido forzados por razones exógenas al proceso generador de la serie, que estamos tratando de elaborar, requieren un trato especial, por lo que debe estimarse su efecto. Algo análogo debe hacerse con el impacto de condiciones meteorológicas especialmente adversas sobre la cosecha de un producto agrícola, cuya producción se pretende modelizar, o con la ocurrencia de una huelga en un sector industrial cuya actividad productiva se quiere prever.

Por tanto, hay muchas instancias en que la estimación de tales efectos por separado del resto del modelo está justificada, A esta estrategia se le denomina Análisis de intervención. Cuando se desconoce la causa de un valor influyente, no procede, en general, llevar a cabo su intervención. Sin embargo, puede intervenirse dicho valor con objeto de que no distorsione la especificación de la estructura estocástica de la serie, para comprobar posteriormente si la estructura que se ha identificado se mantiene al no intervenir el valor influyente, aunque haya que estimar nuevamente el modelo.

Las intervenciones efectuadas sobre un valor influyente son, generalmente, de dos tipos. Si y_{t_0} es un valor influyente, definimos una variable impulso en t_0 como:

$$\xi_{t0}^{i} = \{1 \quad \text{si} \quad t = t_0$$

$$\xi_{t0}^{i} = 0 \quad \text{si} \quad t \neq t_0$$

y una variable escalón en t_0 como:

$$\xi_{t0}^{i} = \{0 \text{ si } t < t_0\}$$

$$\xi_{t0}^{i} = \{1 \text{ si } t > t_0\}$$

El análisis de intervención consiste en introducir una de estas variables en el modelo univariante, y estimar su coeficiente. Se incluye una variable impulso cuando el valor influyente se ha producido tan sólo en t_0. Una variable escalón es más adecuada cuando los valores de la serie (o de sus tasas de cambio) son sistemáticamente mayores o menores después de t_0 que antes de dicho instante o, dicho de otro modo, cuando se ha producido un cambio permanente en el valor medio de la variable, a diferencia del impulso, que corresponde a un cambio meramente transitorio. Cuando la media ha aumentado, el coeficiente asociado a ξ_{t0}^{i} será positivo, siendo negativo en caso contrario.

2.1.12. MODELOS DE FUNCION DE TRANSFERENCIA

La representación ARIMA univariante de una serie temporal correspondiente a una variable Y puede generalizarse para incorporar otras variables X como explicativas. El modelo resultante se conoce como función de transferencia, con las variables X como input e Y como output. Centramos nuestra presentación en el caso de un solo input, pues la extensión a múltiples inputs, tanto en términos de especificación como de estimación, es inmediata.

Conviene especificar un modelo de función de transferencia cuando:

1. Se espera qué la relación entre input y output, tanto a través de sus componentes regulares como estacionales, tenga una estructura dinámica suficientemente rica como para que la representación econométrica habitual requiriese de un número elevado de parámetros.

2. Se cree que una representación adecuada de la estructura estocástica del término de error resultante de la relación de X hacia Y precisa de un modelamiento ARIMA, pues no estaría suficientemente recogida por los sencillos esquemas de autocorrelación utilizados en el modelo lineal general.

Un modelo de función de transferencia tiene varias ventajas adicionales:

1. Permite una representación parsimoniosa, es decir, basada en un número reducido de coeficientes, de relaciones dinámicas, incluso si son muy complicadas.

2. Dispone de una estrategia sencilla y gradual para la especificación de un modelo dinámico de relación que capture adecuadamente el efecto del input sobre el output.

3. Proporciona instrumentos para comprobar que se utilizan como input y output variables que tienen análogas características de estacionariedad, de modo que los residuos del modelo son estacionarios.

Esta propiedad, crucial para justificar el análisis de inferencia que con el modelo estimado pudiera llevarse a cabo, ha generado una considerable cantidad de estudios recientes, destinados a generar procedimientos de contraste de estacionariedad, así como de especificación, ambos en el contexto del modelo econométrico lineal.

La representación genérica del modelo de función de transferencia con un input es:

$$Y_t = v(L)X_t + N_1 = (v_0 + v_1 L + v_2 L^2 + ...)L^b X_t + N_t =$$

$$= (w(L))/(\delta(L))X_{t-b} + N_t = \delta^{-1}(L)w(L)X_{t-b} + N_t =$$

$$= (w_0 - w_1 L - \cdots - w_s L^s)/(1 - \delta_1 L - \cdots - \delta_r L^T)X_{t-b} + N_t$$

donde, al igual que en secciones previas, L denota el operador de retardos.

Los polinomios δ(L) y w(L) se denominan autorregresivo y de medias móviles, respectivamente, mientras que N_t, es la perturbación del modelo. El polinomio v(L) se denomina función de respuesta al impulso, pues sus sucesivos coeficientes describen el efecto que, a través del tiempo, tendría sobre Y un impulso (es decir, un cambio puramente transitorio) en la variable X. El parámetro b, (b > O) que aparece como subíndice en la variable X se denomina tiempo muerto de la relación, y denota el número de períodos que deben transcurrir para que la variación en X comience a dejarse sentir sobre Y.

Por muy rico que sea el efecto de X sobre Y, tanto en intensidad temporal como en duración, es posible recogerlo adecuadamente por medio del polinomio v(L), consiguiendo además que el término de error o perturbación N_t N, sea estacionario. Para ello, puede ser preciso, en ocasiones, utilizar un polinomio v(L) de orden elevado, quizá infinito, pero su descomposición como cociente de dos polinomios de orden finito, incluso muy reducido, como aparece en [2.1.19], permite la representación parsimoniosa de la relación.

A su vez, la perturbación del modelo admitirá una representación univariante:

$$N_t = \psi(L)a_t = (1 + \psi_1 L + \psi_2 L^2 + ...)a_t = (\theta(L)\Theta(L))/(\phi(L)\Phi(L))a_t$$

donde las mayúsculas hacen referencia - a - la - naturaleza estacional de los polinomios respectivos, y donde los polinomios autorregresivos ϕ(L) y Φ(L) pueden contener raíces unitarias, aunque no raíces de módulo superior a la unidad, y a_t, es un ruido blanco gaussiano, es decir, con distribución Normal.

La acumulación sucesiva de los coeficientes de la función de respuesta al impulso genera la función de respuesta al escalón, que proporciona el efecto gradual que sobre Y tendría una desviación permanente en el valor de X. Si la perturbación N_t, es estacionaria, la función de respuesta al escalón es acotada; su límite se denomina ganancia de la relación, y mide el efecto a largo plazo que sobre Y_t, tiene una variación en X_t,. Puede calcularse mediante

$$g = \omega(1) / \delta(1)$$

un instrumento básico en la especificación del modelo de función de transferencia es la función de correlación cruzada (en lo sucesivo, FCC), similar a las funciones de autocorrelación, y definida como cociente entre la covarianza entre X e Y, a distintos retardos y en ambas direcciones, por el producto de sus desviaciones típicas:

$$p_{xy}(k) = Y_{xy}(k) / s_x s_y$$

donde:

$$Y_{xy}(k) = 1/n \sum_{t=1}^{T-k} (x_t - \bar{x})(y_{r+k} - \bar{y}), \qquad \text{k=0,1,2……}$$

$$Y_{xy}(k) = 1/n \sum_{t=k+1}^{T} (x_{t-k} - \bar{x})(y_t - \bar{y}), \qquad \text{k=0, -1, -2…}$$

A diferencia de las FAS y FAP de una serie, no es simétrica, y su valor central ρ_{xy} (O) es el coeficiente de correlación habitual entre ambas variables.

Si las dos variables X e Y no estuviesen correlacionadas entre sí y una de ellas fuese ruido blanco, por ejemplo $Y_t = a_t$, se tendría: -

$$Cov\left[p_{xa}(k), p_{xa}(k+s)\right] \cong p_{xx}(s)/T - k$$

$$Var\left[p_{xa}(k)\right] \cong 1/T - k$$

de modo que:

$$Corr\left[p_{xa}(k), p_{xa}(k+s)\right] \cong p_{xx}(s)$$

En tales condiciones, el inverso del número de grados de libertad da la varianza de cada valor de la FCC, que reproduciría la estructura de la FAS de , X_t. Ello ocurrirá a pesar de la ausencia de correlación cruzada entre ambas variables. Por consiguiente, sólo puede esperarse una FCC con valores nulos fuera de la correlación contemporánea, $p_{xy}(0)$ si ambas series fuesen ruido blanco.

Las expresiones anteriores muestran que los sucesivos valores de la FCC están generalmente correlacionados entre sí, lo que puede dificultar la identificación de la estructura dinámica de la relación entre X e Y. Lo que acabamos de ver es que la FCC tenderá a sugerir más relación entre las variables de la que realmente existe.

2.1.12. a. Identificación del modelo de función de transferencia

La identificación del modelo de transferencia consiste en obtener valores aproximados de los coeficientes de la función de respuesta al impulso v(L) de modo que puedan utilizarse para inferir los órdenes r y s de los polinomios $\delta(L)yw(L)$ y en [2.1.19], así como el tiempo muerto b. Identificados r y s, los coeficientes de v(L) pueden utilizarse asimismo para obtener preestimaciones de los coeficientes $\delta(L)yw(L)$ en y

Si escribimos [2.1.19] en la forma

$$(1-\delta,L-\delta_2L^2-\cdots-\delta_rL^r)(v_0+v_1L+v_2L^2+\cdots)=$$

$$=(w_0-w_1L-w_2L^2-\cdots-w_sL^s)L^b \quad [2.1.20]$$

se tiene el sistema:

$$v_j = 0 \qquad \text{para} \quad j<b$$

$$v_j = \delta_1 v_{j-1} + \partial_2 v_{j-2} + ... + \partial_r v_{j-r} + w_0 \qquad \text{para} \quad j=b$$

$$v_j = \delta_1 v_{j-1} + \partial_2 v_{j-2} + ... + \partial_r v_{j-r} - w_{j-b} \qquad \text{para } j=b+1,\ldots,b+s$$

$$v_j = \delta_1 v_{j-1} + \partial_2 v_{j-2} + ... + \partial_r v_{j-r} \qquad \text{para } j>b+s$$

de modo que los coeficientes de la función de respuesta al impulso tienen la siguiente estructura:

a) Los primeros b coeficientes son nulos, lo que permite identificar el tiempo muerto b: $v_0 = v_1 = ... = v_{b-1} = 0$

b) Desde el coeficiente v_b hasta el coeficiente v_{b+s} no se aprecia ninguna regla de formación.

c) Los coeficientes v_j para $j \geq b+s+1$ se comportan de acuerdo con la ecuación en diferencias de orden r que aparece en [2.1.21], tomando como valores iniciales $v_{b+s`}, v_{b+s-1}, ..., v_{b+s-r+1}$

Por ejemplo, en la función de transferencia:

$$y_t = w_0 /(1 - \delta_1 L - \delta_2 L^2)(X_{t-1})$$

todos los coeficientes de la función de respuesta al impulso obedecen a una ecuación en diferencias de orden 2; lamentablemente, ello puede generar

múltiples configuraciones, aunque una típica sería sinusoidal. - -

Por el contrario, en el modelo $y_t = w_0 - w_1 L - w_2 L^2 / 1 - \delta_1 L(X_{t-1})$

se tendría un primer coeficiente v_0 nulo, seguido de **v_1** y **v_2,** que no obedecerían a ninguna regla, mientras que los sucesivos responderían a una ecuación en diferencias de orden 1, por lo que decaerían exponencialmente, aunque quizá alternando en signo.

Si las variables que se pretende relacionar no son estacionarias, su función de correlación cruzada, así como sus FAS y FAP, no decaerá rápidamente hacia cero. En tal caso, es preciso transformarlas mediante diferencias para lograr estacionariedad, lo que denotamos en lo sucesivo por

$$y_t = \nabla^d Y_t, x_1 = \nabla^d X_{t,}, n_t = \nabla^d N_t$$

que serían estacionarias. Si es preciso tomar asimismo diferencias de orden estacional, así como quizá tomar logaritmos para atenuar la heteros-cedasticidad, todo ello estaría incorporado en nuestra notación y_t, x_t, n_t: La función de transferencia transformada resulta:

$$y_t = v_0 x_t + v_1 x_{t-1} + v_2 x_{t-2} + ... + n_t \quad [2.1.22)$$

que conserva los mismos coeficientes de la original. Supongamos que esta función de transferencia es tal que v_j para **j>** q. Si premultiplicamos en

[2.1.22] por x_{t-k} para sucesivos valores k ~ O y tomamos esperanzas se tiene:

$$\begin{pmatrix} y_{xy}(0) \\ y_{xy}(1) \\ y_{xy}(q) \end{pmatrix} \quad \begin{pmatrix} y_{xx}(0)\ y_{xx}(1)\ .....\ y_{xx}(q) \\ y_{xx}(1)\ y_{xx}(0)\ .....\ y_{xx}(q-1) \\ y_{xx}(q)\ y_{xx}(q-1)\ .....\ y_{xx}(0) \end{pmatrix} \quad \begin{pmatrix} v_0 \\ v_0 \\ v_q \end{pmatrix}$$

de modo que si se dispusiese de las estimaciones muestrales de la FCC entre x e y, lo que siempre es posible, podríamos resolver el sistema [2.1.23] para obtener estimaciones de los coeficientes de la función de respuesta al impulso $v_0, v_1 .... v_n$

2.1.12.b. Identificación con preblanqueo

El proceso de identificación de la función de transferencia se simplifica notablemente si se filtran previamente ambas variables de manera adecuada. Su pongamos que la variable x_t siendo estacionaria, admite la representación ARMA: $\phi_x(L)\theta_x^{-1}(L)x_t = a_t$

donde α_t ; es ruido blanco y $\phi_x(L)$ y $\theta_x(L)$ son los polinomios autorregresivos y de medias móviles de dicho modelo ARMA. Una vez

estimado [2.1.24], obtendríamos los residuos $\hat{\alpha}_t$ que no son sino el resultado de filtrar la variable x_t , por su modelo ARMA. Supongamos que filtramos el output y_t, con el mismo modelo ARMA del input que acabamos de estimar, obteniendo:

$$\beta_t = \phi_x(L)\theta_x^{-1}(L)y_t$$

Las variables así filtradas (no cabe esperar que b_t, sea ruido blanco) satisfacen la relación:

$$\beta_1 = v(L)a_t + \varepsilon_t$$

donde:

$$\varepsilon_t = \phi_x(L)\theta_x^{-1}(L)n_t$$

pudiendo observarse que la función de respuesta al impulso en [2.1.25] es idéntica a la original. Si multiplicamos por α_{t-k} y tomamos esperanzas, se tiene:

$$\gamma_{a\beta}(k) = v_k\sigma_a^2$$

donde $\gamma_{\alpha\beta}$ denota la covarianza entre α_t y β_t en el retardo k, por lo que:

$$v_k = y_{a\beta}(k)/\sigma_a^2 = p_{a\beta}(k)/\sigma_a, k = 0,1,2...$$

de modo que tras preblanquear input y output con el modelo del input, la

función de correlación cruzada es proporcional a la función de respuesta al impulso. El efecto de preblanquear es transformar el sistema de ecuaciones [2.1.23] en un conjunto de ecuaciones como [2.1.26], que pueden resolverse por separado para obtener preestimaciones de la función de respuesta al impulso a partir de las desviaciones típicas y de la función de correlación cruzada muestra! de α y β

2.1.12.c. Identificación de un modelo para el ruido

Una vez identificada una función de respuesta al impulso, y habiendo obtenido estimaciones para sus coeficientes $\bar{v}(L)$ podemos estimar el ruido: $\hat{n}_t = y_t = -\hat{v}(L)x_t$

Alternativamente, suele comenzarse especificando para n_t, el muestreo univariante de y_t, lo que sería correcto si no existiese relación entre y_t, y x_t en cuyo caso, la función de transferencia sería nula. Al ir estimando coeficientes de v(L), el modelo de n_t, se hace más sencillo.

Ambas opciones se basan en modelos estimados previamente, por lo que todo lo que hacemos a continuación tiene un valor tan sólo aproximado. En definitiva, ha de ser el rigor del analista el que debe conducir a la especifica-

ción de un modelo univariante, como los estudiados en las secciones previas, para los residuos resultantes del modelo de función de transferencia.

Al especificar un modelo de transferencia, Box y Jenkins sugieren: -

1. Que un modelo con estructuras AR o MA de orden 1 ó 2 (además del posible tiempo muerto) será normalmente suficiente, si bien esto debe tenerse en cuenta tanto para la componente regular como para la estacional.
2. Que, aunque la estimación del modelo de transferencia es eficiente sólo si el modelo especificado es correcto, las estimaciones de los valores v_k de la función de respuesta al impulso son útiles en la identificación de la estructura
dinámica de relación.
3. Que tampoco se gana mucho con tratar de obtener estimaciones eficientes de los v_k pues - son precisos muchos coeficientes de tal tipo para resumir el cociente de polinomios $\delta(L)/\omega(L)$**-** Recíprocamente, el objetivo último del analista es obtener estimaciones precisas de estos dos polinomios, y las estimaciones que ellas implican para los coeficientes v_k están altamente correlacionadas entre sí, y tienen una elevada varianza.

2.1.12.d. Estimación de un modelo de función de transferencia

Una vez identificado el modelo de transferencia, sin estacionalidad:

$$y_t = \delta^{-1}(L)\omega(L)x_{t-b} + \phi^{-1}(L)\theta(L)a_t$$

y dados unos valores iniciales x_0, y_0, a_0 podemos obtener los residuos sucesivos: $a_t(b, \delta, \omega, \phi, \theta / x_0, y_0, a_0$ a partir de t_0 = máx {r, s + b} + 1 y, bajo el supuesto uc i~ormaiiuau, ratar de minimizar su suma de cuadrados. Como ejemplo, consideremos la función de transferencia:

$$y_t = \omega_0 + \omega_1 L + \omega_2 L^2 \,/1 - \delta L(x_t) + 1/1 - pL(a_t)$$

en la que máx (r,s+b)+1=3 , y donde $n_t = a_t /(1 - pL)$ en la que podemos hacer:

$$n_3 = y_3 - \delta y_2 - w_0 x_3 - w_1 x_2 - w_2 x_1 + \delta n_2$$

$$n_4 = y_4 - \delta y_3 - w_0 x_4 - w_1 x_3 - w_2 x_2 + \delta n_3$$

comenzando de n_2 = O para después obtener las innovaciones a_t, a partir de t =4, por medio de:

$$a_4 = n_4 - pn_3$$

$$a_5 = n_5 - pn_4$$

y formar, por último, la suma residual:

$$S^2(b, \delta, \omega, p) \equiv \sum_4^T a_t^2 (a, \delta, \omega, p / x_{0,} y_0, a_0$$

donde las primeras innovaciones se han tomado iguales a su esperanza incondicional, que es cero.

La expresión genérica de cada a, es:

$$a_t = y_t - \delta y_{t-1} - \omega_0 x_1 - \omega_1 x_{t-1} - \omega_2 x_{t-2} - + \delta n_{t-1} - p\left[y_{t-1} - \delta y_{t-2} \omega_0 x_{t-1} - \omega_1 x_{t-2} - \omega_2 x_{t-3} + \delta n_{t-2}\right]$$

con vector gradiente:

$$\partial a_t / \partial \delta = -y_{t-1} + p y_{t-2} + n_{t-1} - p n_{t-2}$$

$$\partial a_t / \partial \omega_0 = -x_t + p x_{t-1}$$

$$\partial a_t / \partial \omega_1 = -x_{t-1} + p x_{t-2}$$

$$\partial a_t / \partial \omega_2 = -x_{t-2} + p x_{t-3}$$

$$\partial a_t / \partial \rho = -(y_{t-1} - \delta y_{t-2} - \omega_0 x_{t-1} - \omega_1 x_{t-2} - \omega_2 x_{t-3} + \delta n_{t-2}$$

por lo que podemos comenzar un algoritmo iterativo del tipo Gauss-Newton a partir de estimaciones iniciales $\hat{\theta} = (\hat{\delta}, \hat{w}_0, \hat{w}_1, \hat{w}_2 \hat{p})$. Como allí se vio, un modo de proceder consiste en obtener series temporales para cada una de las componentes del gradiente, así como para los propios residuos a_t, utilizando las estimaciones iniciales de los parámetros, para estimar la

regresión:

$$d_t = (\hat{\delta} - \delta)\partial a_t(\hat{\theta})/\partial\delta + (\hat{w}_0 - w_0)\partial a_t(\hat{\theta})/\partial w_0 + (\hat{w}_1 - w_1)\partial a_t(\hat{\theta})/\partial w_1 +$$

$$+ (\hat{w}_2 - w_2)\partial a_t(\hat{\theta})/\partial w_2 + (\hat{p} - p)\partial a_t(\hat{\theta})/\partial p + a_t$$

que se repite, tomando en cada iteración las últimas estimaciones como iniciales, hasta lograr la convergencia. La aparición del vector $\hat{\theta}$ hace referencia a que las componentes del gradiente de la función a, están evaluadas en las preestimaciones. Este procedimiento está sujeto a todas las consideraciones que se hicieron en relación con la convergencia de un algoritmo numérico. Lograda la convergencia, la matriz de - covarianzas de las estimaciones se aproxima por el producto $\hat{\sigma}_a^2(\nabla a \nabla a`)^{-1}$ donde σ_a^2 se estima mediante el cociente del valor alcanzado por la suma residual en la última iteración, y el número de observaciones, o de grados de libertad, si se prefiere.

2.1.12.e. Diagnóstico del modelo de función de transferencia

Uno de los análisis que debe hacerse de toda función de transferencia estimada es la búsqueda de posibles factores comunes que pudiesen

quedar enmascarados en los polinomios AR y MA que la componen. Nótese que el modelo de función de transferencia, al igual que todo modelo lineal, es único excepto por operadores factoriales, posiblemente dinámicos. Es decir, si el «verdadero» modelo de transferencia es: -

$$Y_1 = \delta^{-1}(L)\omega(L)X_{t-b} + \phi^{-1}(L)\theta(L)a_t$$

También es válido el modelo: -

$$D(L)Y_1 = D(L)\delta^{-1}(L)\omega(L)X_{t-b} + D(L)\phi^{-1}(L)\theta(L)a_t$$

donde D(*L*) es un polinomio cualquiera en el operador de retardos, aunque el analista preferirá el primero por su sencillez; en la práctica, el menor número de parámetros se traducirá en una mayor eficiencia en la estimación, así como en una ausencia de correlaciones entre sus valores estimados. El objetivo del investigador ha de ser siempre lograr un modelo tan sencillo como sea posible, comenzando siempre por especificaciones sencillas, para complicarlas posteriormente, sólo si es preciso. En todo caso, las correlaciones entre los parámetros estimados deben ser reducidas, pues valores elevados indican sobre parametrización.

Supongamos que el modelo correcto es $y_t = v(L)x_t + \psi(L)a_t$ pero el investigador, incorrectamente, especifica $y_t = v_0(L)x_t + \psi_0(L)a_{0t}$ que, una vez estimado, genera unos residuos que denotamos por a_{0t} . estos quedan defini-

dos por:

$$a_{0t} = \psi_0^{-1}(L)\left[v(L) - v_0(L)\right]x_t + \psi_0^{-1}(L)\psi(L)a_t$$

donde puede apreciarse que la mala especificación puede hacer que: a) los residuos tengan autocorrelación, y b) estén correlacionados con las variables, x_t, y por tanto, con el ruido blanco α_t ; que genera las x_t.

Caso 1: El modelo de transferencia se especifica correctamente, pero el modelo del ruido incorrectamente: $v_0(L) = v(L)$ **,** pero $\psi_0(L) \neq \psi(L)$**.** En tal caso: $a_0^{-1} = \psi_0^{-1}(L) \neq \psi(L)a_t$ por lo que a) la función de correlación cruzada entre los residuos y el input del modelo debe ser no significativamente distinta de cero, pero b) los residuos presentan autocorrelación.

Caso 2: El modelo de transferencia se especifica incorrectamente. En este caso se tiene $a_0^{-1} = \psi_0^{-1}(L)\left[v(L) - v_0(L)\right]x_t + a_t$ por lo que se tiene tanto autocorrelación de los residuos, como una función de correlación cruzada significativamente distinta de cero con el input.

La conveniencia de preblanquear el input antes de obtener su función de correlación cruzada con los residuos proviene de que, como vimos al comienzo de esta sección, los sucesivos valores de dicha función de correlación cruzada están correlacionados, lo que no ocurre con la función

de correlación cruzada entre el input preblanqueado y los residuos, si éstos son efectivamente ruido blanco.

Pero además, este mismo análisis sugiere el tipo de modificaciones que deben introducirse en la función de transferencia previamente estimada para obtener una mejor especificación. En efecto, considerando el modelo pre-blanqueado: $\beta_t = v(L)\alpha_t + \varepsilon_t$ y si denotamos por ε_{0t}, los residuos del modelo incorrecto,se tiene:

$$\varepsilon_{0t} = \left[v(L) - v_0(L)\right]\alpha_t + \varepsilon_t$$

se tiene:

$$v_k - v_{0k} = \rho_{a,\varepsilon 0}(k).\sigma_{\varepsilon 0} / \sigma_a, k = 0,1,2,.....$$

de modo que la función de correlación cruzada entre los residuos y el input preblanqueado mide la discrepancia entre la función de respuesta al impulso «verdadera» y la estimada.

Por último, puesto que es necesario para una correcta especificación de la función de transferencia que los residuos estén libres de autocorrelación, puede utilizarse los estadísticos de Box-Pierce o de Ljung-Box. En este caso, $Q_g = T\sum_0^g \hat{r}_a^1(k)$ se distribuye como una $X^2_{g+1-(r+s-1)}$, donde r y s son los órdenes de los polinomios AR y MA de la función de transferencia, y no dependen del modelo del ruido.

2.2 ANALISIS DE COMPONENTES PRINCIPALES

En la investigación de mercado se desea obtener los factores que tienen la mayor influencia en la decisión de compra, para esto es importante reducir el número de factores y el tratamiento de componentes principales es importante. A continuación detallaremos la teoría utilizada correspondiente al Análisis de Componentes Principales.

2.2.1.- INTRODUCCIÓN.

A lo largo de todo este desarrollo histórico se han planteado algunos problemas de fondo que han dado lugar a distintas propuestas de solución, los aspectos más polémicos, entre otros, han sido:

- La estimación de las comunalidades.
(Comunalidad: grupo de personas que ocupan una o varias zonas geográficas que tiene similitud en sus costumbres, organización y gobierno.)
- Los métodos de extracción de factores.
- El Nº de factores a retener
- Los métodos de rotación de los factores

Se han propuesto múltiples métodos para la extracción de factores que conducían a soluciones diferentes según el método que se adoptase. Las respuestas han sido distintas según las diversas tendencias. Algunos autores consideran el ACP como una etapa del Análisis Factorial y otros lo consideran como técnicas diferentes.

Lo que parece claro es que ambos métodos parten de una misma premisas. Un espacio R^k en el cual se sitúa una nube de puntos N cada uno de ellos con una masa, y en los cuales se define una métrica, se calcula la inercia total de la nube y se determinan los ejes de inercia.

Los “inputs” de un análisis factorial, en todos los casos, son los siguientes: El espacio. Los puntos, los pesos que afectan a los puntos, la métrica. Los “outputs” son los ejes de inercia, las coordenadas de los puntos sobre los ejes.

De un método a otro solo varia los “inputs”, la definición de los puntos, los pesos y la métrica. Según se considere un análisis u otro. Pero una vez dados los “inputs” lo esencial de la técnica es común a todos los métodos

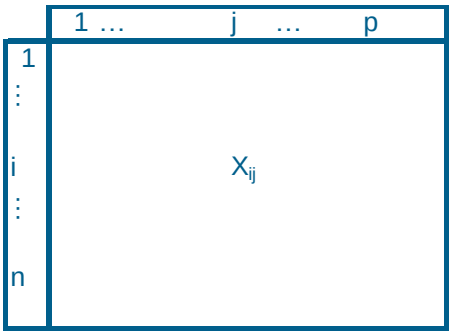
Desde los años 60, la evolución de estas técnicas ha estado sometida a la

complejidad de su cálculo debido a la gran masa de datos a analizar. Desde la aparición de los sistemas informáticos esta gran masa de datos ha sido posible manejarla en su conjunto y las técnicas de análisis multivariante han supuesto una revolución en la investigación. Hoy en día, se dispone de gran cantidad información que da pie a estudiar fenómenos (calidad de vida, nivel socioeconómico, grado de bienestar, etc...) que no pueden medirse directamente, sino que son el resultado del estudio y análisis un conjunto de variables relacionadas.

El Análisis Multivariante puede considerarse como un conjunto de técnicas o métodos científicos que permiten tratar matrices de grandes dimensiones. Dentro de este análisis podemos considerar diferentes técnicas, un grupo de estas parte de matrices en que los datos corresponden: por filas a individuos y por columnas a variables otras parten de matrices cuyos datos tanto de filas como columnas, se refieren a variables. Entre las primeras podemos destacar al Análisis de Componentes Principales (A.C.P.) y el Análisis Factorial (A.F.), entre las segundas tenemos, entre otras, al Análisis de Correspondencias.

En ACP los datos se aplican a tablas bidimensionales que cruzan individuos y variables cuantitativas. Las filas representaran a los individuos y las columnas a las variables.

Gráfico 2.2.1.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Estructura de datos para un ACP



Autor: José Aguayo E. Fuente: ACP por Salvador C. Arroyo de la Universidad de Valencia.

Desde un punto de vista estadístico muy general el objetivo prioritario de un análisis multivariante será reducir la dimensión original de un conjunto de *p* variables, a un conjunto menor de *m* variables para lograr una mayor interpretabilidad de la información. Lo que se pretende es reducir el número de variables a utilizar manteniendo el máximo de información sin redundancias, reduciendo la dimensionalidad del espacio original de manera que estas nuevas variables sintéticas expliquen la máxima variabilidad total de las variables originales (con la distorsión mínima de la información). Estas m nuevas variables serán variables no observables o latentes, que se determinaran como veremos mediante combinación lineal de las variables

originales.

Centrándonos en el Análisis de Componentes Principales, el análisis sirve para reducir un conjunto de variables originales p relacionadas a un numero menor m de nuevas variables, que llamaremos componentes principales, independientes entre si. Con ello lo que usualmente se consigue es agrupar las variables originales en subconjuntos de variables que están relacionadas entre sí y no están relacionadas con las variables de los otros subconjuntos. Este subconjunto de variables relacionadas entre si que se constituyen como combinación lineal tienen la propiedad de explicar parte de la varianza de las variables originales. Así el objetivo del ACP será obtener el mínimo número de componentes que expliquen en su conjunto la máxima varianza de las variables originales.

En general, existen dos enfoques para utilizar el ACP: el exploratorio y el confirmatorio. El primero, el investigador parte de una información en la que desconoce las interrelaciones de las variables originales, como se organizan y por tanto no tiene una idea clara de lo que puede encontrar. En el segundo, lo que pretende es corroborar la existencia de determinadas agrupaciones de las variables originales que conformen una idea prefijada.

Se va a integrar ambas corrientes de manera que se pueda plasmar de la manera más sencilla posible ambas metodologías. Así podemos realizar dos análisis duales: uno en el espacio de los individuos y otro en el de las variables.

En un primer estadio del análisis podemos preguntarnos por las semejanzas o no de los individuos así como las relaciones posibles existentes entre las variables. De esta forma podremos interpretar las semejanzas de los individuos, la existencia de grupos homogéneos de individuos llegando a poner en evidencia una determinada tipología de estos. Desde el punto de vista de las variables podremos explicar sus relaciones preguntándonos que variables son las que están relacionadas positiva o negativamente entre si y también si podemos establecer una tipología de las variables. Quedaría buscar si existe una relación entre ambas tipologías, además de relacionar cada uno de los individuos con el conjunto de las variables originales y las nuevas variables (componentes).

2.2.2.- ESTRUCTURA DE DATOS: estandarización escala de medidas.

Una de las cosas que tenemos que tener en cuenta en general es que las

matrices de datos X no estarán medidas en una escala de medida única, por lo que es conveniente antes de efectuar cualquier análisis asegurarse de la homogeneidad de los datos, ya que la técnica descansa en la geometría analítica y utiliza la “distancia” como instrumento de interpretación de las semejanzas anteriormente expuestas, de esta forma tendrá sentido interpretar la “distancia” entre las filas (individuos) y las columnas (variables).

Recordemos que estandarizar o tipificar una variable requiere una serie de pasos como son: en primer lugar centrado de la variable, esto es, restarle su media y en segundo lugar dividir por su desviación y dado que la desviación esta definida en la misma unidad de medida que la variable original, las nuevas variables estandarizadas no dependen de dicha unidad de medida y así pueden compararse variables originales medidas con diferentes escalas. Es decir: centrar es restar a cada valor numérico (dato) la media de la variable correspondiente. La tabla que resulta tendrá entonces como término general (x_{ij})

$$x_{ij} = (x_{ij} - \bar{x})$$

Esta transformación no tiene influencia alguna sobre las definiciones de semejanza entre los individuos y de relación entre las variables y por tanto este centrado no modifica la unidad de medida original. La manera más

sencilla de eliminar las unidades de medida es tipificar los datos, es decir, dividir cada dato centrado o no por la desviación típica correspondiente. En estadística se denomina variable tipificada, estandarizada o reducida a la que esta centrada, su media es cero y esta dividida por su desviación (σ_j), donde su término general seria:

$$\left(\frac{x_{ij} - \overline{x_j}}{\sigma_j} \right)$$

De esta manera, todas las variables presentan entonces la misma variabilidad y por ello tendrán la misma influencia en el cálculo de las distancias - semejanzas entre los individuos.

2.2.3.- AJUSTE DE LA NUBE DE INDIVIDUOS Y VARIABLES: Análisis en R^P y R^N

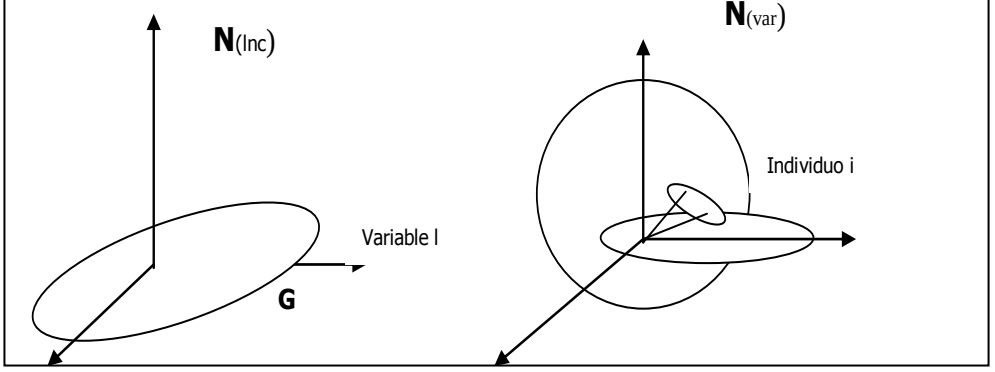
Recordemos que al hablar en la introducción definíamos la matriz de datos como una matriz que por filas se identifica con los individuos y por columnas con las variables.

Partimos de un conjunto de n individuos i =1... n sobre el que se observan p variables j=1....p

$$x = \begin{bmatrix} x_{11} & x_{21} & . & x_{1p} \\ x_{21} & x_{22} & . & x_{2p} \\ . & . & . & . \\ x_{n1} & x_{n2} & . & x_{np} \end{bmatrix}$$

Cada individuo ó fila es un vector de p componentes donde cada uno de ellos esta asociado a una variable. Bajo este punto de vista se puede representar a cada individuo como un punto o vector en el espacio p-dimensional (R^p) en el que cada dimensión (que representa a un individuo) esta referida a una variable.

Gráfico 2.2.2.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Representación de nubes de individuos



Autor: José Aguayo E. Fuente: ACP por Salvador C. Arroyo de la Universidad de Valencia.

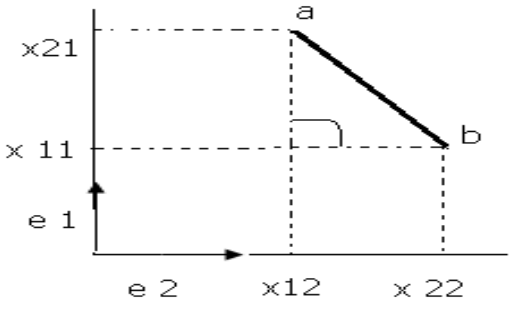
El conjunto de individuos constituye la nube que podemos identificar como N_{ind} en la que los datos han sido centrados de manera que su centro coincide con

el origen de los ejes como consecuencia de haber sido previamente centrada, a ese centro de coordenada se le llama centro de gravedad y se le representa por G_{ind} que como veremos cumple la particularidad de representar al individuo medio. De esta manera el análisis de componentes principales le llamaremos normado.

También veremos que la posición media de los n individuos respecto a las p variables viene dada por el vector medias $\tilde{x}$.

En cualquier espacio multidimensional puede definirse una “distancia” entre cada par de individuos. La conceptualización matemática de la idea de distancia debe cumplir unos axiomas que vamos a obviar, pero debe quedar muy claro que pueden definirse diferentes tipos de distancias y todas ellas deben cumplir su axiomática de definición. La distancia mas intuitiva entre dos puntos es la euclidea que viene dada por la diagonal (o su cuadrado) del triángulo rectángulo que se puede construir entre dichos puntos y sus proyecciones perpendiculares a los ejes.

Gráfico 2.2.3.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Proyecciones



Autor: José Aguayo E. Fuente: ACP por Salvador C. Arroyo de la Universidad de Valencia**.**

os vectores a y b se descomponen sobre la base e1 y e2.

Siendo:

$$\|e_1\| = 1$$
$$\|e_2\| = 1$$
$$< e_1; e_2 >= 0$$

de manera que:

$$a = x_{12}`.e_2 + x_{21}.e_1$$
$$b = x_{22}.e_2 + x_{11}.e_1$$

el cuadrado de la distancia entre los vectores a y b será:

$$d^2(a;b) = (x_{12} - x_{22})^2 + (x_{21} - x_{11})^2$$
$$d^2(a;b) = \|a - b\|^2$$
$$d^2(a;b) = < a - b; a - b >$$

También se puede expresar la distancia como:

$$d^2(a;b) = (a - b)`.M.(a - b)$$

donde **M** es una matriz definida positiva que en el caso de utilizar una estructura euclidea coincide con la matriz identidad **I,** pasando así a definir la métrica del espacio. Podemos definir el producto escalar de dos vectores a y b en el espacio de individuos como:

$$< a; b >_M = a' Mb$$

En nuestro caso la generalización de la distancia euclidea para p variables es inmediata y se considera como distancia euclidea al cuadrado el valor:

$$d^2(i, i`) = \sum_{j=1}^{p} (x_{ij} - x_{i`j})^2$$

En el espacio R^p de las variables, la noción de semejanza entre dos individuos coincide con la distancia euclidea. El conjunto de distancias inter-individuos constituye lo que se llama la forma de la nube N_N (individuos).

2.2.3.a.- Pesos de individuos y variables

Antes de abordar cualquier tipo de análisis conviene tener en cuenta los datos de partida y si les vamos a dar la misma importancia o no. Así cuando existen individuos procedentes de poblaciones con mayor o menor importancia les podemos asignar un “peso” diferente según la población de procedencia. Una de las formas podría ser asignándole un peso proporcional al efectivo “total de elementos” de la susodicha población.

En el caso de que cada uno de los individuos represente una población con mayor o menor importancia, le asignaremos un peso proporcional al efectivo de la población que represente. Así llamaremos p_i al peso asignado al individuo i.

En la mayor parte de los casos por comodidad se toman los pesos de manera que la masa total de estos individuos sea la unidad, asociando a cada individuo el peso 1/N. Por lo tanto tenemos que:

$$\overline{x_{k0}} = \sum_{i=1}^{n} \frac{x_{ik}}{N} = \sum_{i=1}^{n} \frac{1}{N} \cdot x_{ik} = \sum_{i=1}^{n} p_i \cdot x_{ik}$$

$$r_{kh} = \sum_{i=1}^{n} \frac{\frac{\left(x_{ik} - \overline{x_k}\right)\left(x_{ih} - \overline{x_h}\right)}{N}}{\sigma_k \cdot \sigma_h} = \sum_{i=1}^{n} \frac{1}{N}\left(\frac{x_{ik} - \overline{x_k}}{\sigma_k}\right)\left(\frac{x_{ih} - \overline{x_h}}{\sigma_h}\right) =$$

$$= \sum_{i=1}^{n} p_i \left(\frac{x_{ik} - \overline{x_k}}{\sigma_k}\right)\left(\frac{x_{ih} - \overline{x_h}}{\sigma_h}\right)$$

En el caso de las variables y de forma similar, la importancia de unas y otras variables se puede modular utilizando un coeficiente llamado peso de la variable. Así llamamos m_j al peso de la variable j, y la distancia entre los dos individuos i y i' viene definida por:

$$d^2(i, i`) = \sum_{j=1}^{p} m_j (x_{ij} - x_{i`j})^2$$

Vamos a considerar a partir de ahora que:

$$p_i = \frac{1}{N} \rightarrow \forall_i \in N$$
$$m_j = 1 \rightarrow \forall_j \in P$$

Efectuar un análisis de estas distancias supone estudiar la forma de la nube, es decir descubrir una partición de los puntos o direcciones de alargamiento.

De forma similar cuando hablamos de variables cada una de ellas tiene sus componentes asociadas a n filas (individuos). Bajo este punto de vista, se

puede representar cada variable como un vector del espacio vectorial R^N, en el que cada dimensión esta referida a un individuo: por ejemplo la variable j esta representada por el vector simbolizado $\vec{x}$ y cuya componente i-esima es:

$$\frac{x_{ij} - \overline{x_j}}{\sigma_j}$$

El conjunto de los puntos extremos de los vectores que representan las variables constituye la nube N_p (variable).

La distancia entre variables en R^N consiste en afectar a cada dimensión de un coeficiente igual al peso de cada individuo en la nube N^N de R^p. En el caso general en que los pesos coincidan, la distancia utilizada solo difiere de la euclidea usual en el coeficiente 1/N.

En un principio la distancia entre dos variables originales es:

$$d^2(j, j`) = \sum_{i=1}^{n} (x_{ij} - x_{ij`})^2$$

Cuando utilizamos las variables previamente tipificadas, por los motivos aludidos anteriormente tenemos que la distancia será:

$$d^2(j, j`) = \sum_{i=1}^{n} \left[ \left( \frac{x_{ij} - \overline{x_j}}{\sigma_j} \right) - \left( \frac{x_{ij`} - \overline{x_{j`}}}{\sigma_{j`}} \right) \right]^2$$

Si introducimos 1/N estamos utilizando otra métrica diferente de la euclidea en

esa proporción.

Al estar la nube centrada sobre el origen la distancia en esta nueva métrica de una variable al origen será:

$$d^2(j,G) = \frac{1}{N}\sum_{i=1}^{n}\left[\left(\frac{x_{ij} - \overline{x_j}}{\sigma_j}\right) - 0\right]^2 = \frac{1}{N}\sum_{i=1}^{n}\frac{\left(x_{ij} - \overline{x_j}\right)^2}{\sigma_j^2} = \frac{\sigma_j^2}{\sigma_j^2} = 1$$

Con esta distancia, los vectores que representan las variables centradas y tipificadas tienen las siguientes propiedades:

a) Cada vector, que es representación de una variable, tiene como norma la unidad

$$\left\|\overline{k}\right\|^2 = \sum_{i=1}^{n}\frac{1}{N}\left(\frac{x_a - \overline{x_k}}{\sigma_k}\right)^2 = 1 = d^2(j,G)$$

Por ello la nube N_p esta repartida sobre una esfera de radio unidad.

También podemos observar que:

$$d^2(j, j`) = \frac{1}{N}\sum_{i=1}^{n}\left[\frac{\left(x_{ij} - \overline{x_j}\right)}{\sigma_j} - \frac{\left(x_{ij`} - \overline{x_{j`}}\right)}{\sigma_{j`}}\right]^2 =$$

$$= \frac{1}{N}\sum_{i=1}^{n}\left[\frac{\left(x_{ij} - \overline{x_j}\right)^2}{\sigma_j^2} + \frac{\left(x_{ij`} - \overline{x_{j`}}\right)^2}{\sigma_{j`}^2} - 2\frac{\left(x_{ij} - \overline{x_j}\right)\left(x_{ij`} - \overline{x_{j`}}\right)}{\sigma_j \sigma_{j`}}\right] =$$

$$= \frac{\sigma_j^2}{\sigma_j^2} + \frac{\sigma_{j`}^2}{\sigma_{j`}^2} - 2\frac{\sigma_{jj`}}{\sigma_j \sigma_{j`}} = 2 - 2r_{jj`} = 2(1 - r_{jj`})$$

Es decir:

$$d^2(j, j`) = 2(1 - r_{jj`})$$

esto implica que:

$$0 \le d^2\left(j, j`\right) \le 4$$

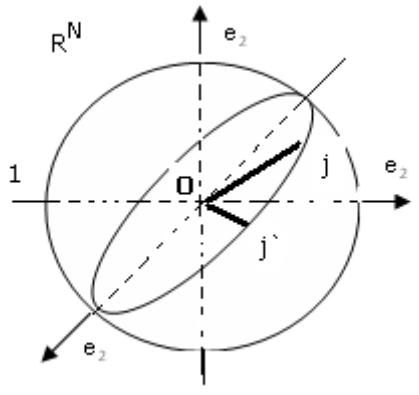
donde r_{jj} el coeficiente de correlación entre las dos variables j y j'. Así las proximidades entre las dos variables se pueden interpretar en términos de correlación. Si la distancia entre las dos variables esta próxima a cero la correlación será prácticamente 1. Si están muy correlacionadas negativamente r_{jj} =-1 la distancia será máxima e igual a 4. Si están incorrelacionadas, el coeficiente de correlación es cero, la distancia es 2 (intermedia).

b) El coseno del ángulo que forman los vectores que están representando a las dos variables j y j' coinciden con el coeficiente de correlación entre ambas. Así:

Gráfico 2.2.4.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Correlación entre variables



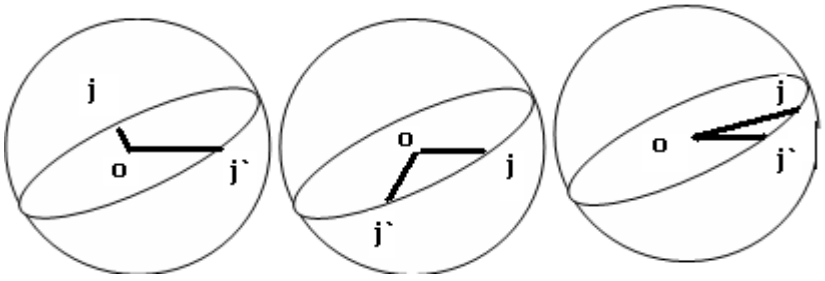
Autor: José Aguayo E. Fuente: ACP por Salvador C. Arroyo de la Universidad de Valencia**.**

$$\cos(j, j`) = < j, j` > = \sum_{i=1}^{n} \frac{1}{N} \left(\frac{x_{ij} - \overline{x_j}}{\sigma_j} \right) \left(\frac{x_{ij} - \overline{x_{j`}}}{\sigma_{j`}} \right) =$$

$$= \sum_{i=1}^{n} \left(\frac{x_{ij} - \overline{x_j}}{\sqrt{N}\sigma_j} \right) \left(\frac{x_{ij} - \overline{x_{j`}}}{\sqrt{N}\sigma_{j`}} \right) = r_{ij`}$$

La interpretación de un coeficiente de correlación como un coseno justifica la métrica elegida.

Gráfico 2.2.5.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Correlación entre variables



Autor: José Aguayo E. Fuente: ACP por Salvador C. Arroyo de la Universidad de Valencia.

$$r_{jj`} \cong -1 \qquad r_{jj`} \cong 0 \qquad r_{jj`} \cong 1$$
$$d(jj`) = 2 \qquad d(jj`) = \sqrt{2} \qquad d(jj`) = 0$$

Al ser la longitud de los vectores que representan a las variables igual que la unidad, la coordenada de la proyección de una variable sobre otra se puede tomar como medida del mutuo coeficiente de correlación por tanto. Efectuar un análisis de coeficientes de correlación entre las variables supone estudiar los ángulos entre los vectores que definen esta nube N_p. Tal estudio es imposible de realizar directamente dada la dimensión de R^N. De ahí el interés del ACP al proporcionar variables sintéticas que constituyen un resumen del conjunto de variables originales.

2.2.3.b.- Centro de gravedad e inercia de una nube de puntos

Consideremos la información proporcionada en una tabla de individuos por variables (matriz X): cada individuo tiene unas características dadas por la fila que le corresponde. Los p datos del individuo configuran un vector ($\vec{x} \equiv x_{i1}$…. x_{ip}) que se representa como un punto en el espacio R^p Los n individuos forman una nube de n puntos en R^p.

Cuando solamente hay dos variables (p=2) esa nube es más fácil de interpretar. Así:

Gráfico 2.2.6.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Gráfico de nube de puntos

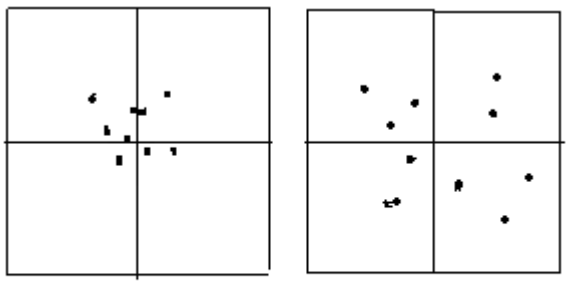


Grafico A: poca inercia

Grafico B: mucha inercia

Autor: José Aguayo E. Fuente: ACP por Salvador C. Arroyo de la Universidad de Valencia.

Un conjunto de individuos puede caracterizarse por su gravedad y por su inercia.

Como dijimos el centro de gravedad marcado en los gráficos como punto $\bar{x}$, es el vector medias:

$$\bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_p \end{bmatrix}$$

Es el punto que señala la posición central de la nube, caracterizando al individuo promedio respecto a las p variables.

La inercia de una nube de puntos es una medida sintética de dispersión, se define como la suma para todos los puntos del producto de sus masas por los cuadrados de sus distancias al centro de gravedad.

$$inercia = \sum_{i=1}^{n} m_i . d^2 \left(x_i \bar{x} \right)$$

Según esta medida sintética de dispersión el gráfico A tiene poca inercia, ya que los individuos son muy homogéneos, situándose cerca del centro de gravedad. Al contrario el gráfico

Cuando se adopta la distancia euclidea ordinaria, la inercia de una nube de

puntos es la suma de las varianzas de las p variables.

$$d^2\left(x_i, \bar{x}\right) = \sum_{j=1}^{p}\left(x_{ij} - \overline{x_j}\right)^2$$

y obtenemos:

$$inercia = \sum_{i=1}^{n} m_i \sum_{j=1}^{p}\left(x_{ij} - \overline{x_j}\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} m_i\left(x_{ij} - \overline{x_j}\right)^2$$

$$como..m_i = \frac{1}{N} \rightarrow \forall i$$

tenemos que:

$$\sum_{I=1}^{N} \frac{1}{N}\left(x_{ij} - \bar{x}_j\right)^2 = \sigma_j^2$$

Luego la inercia es la varianza:

$$INERCIA = \sum_{J=1}^{N} \sigma_J^2$$

Luego la inercia de la nube formada por los n individuos se calcula sumando los valores de la diagonal principal de la matriz de covarianzas 5

inercia = Traza de S

La matriz de covarianzas S será:

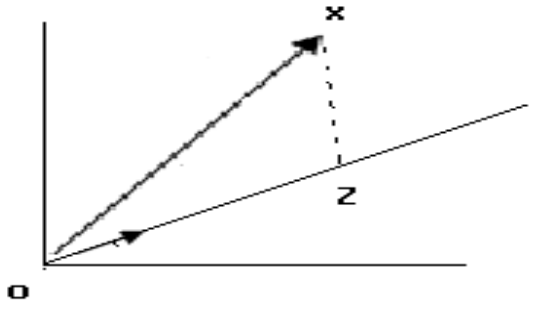
$$S = \frac{1}{N}\begin{bmatrix} x_{11} & x_{21} & . & x_{n1} \\ x_{12} & x_{22} & . & x_{n2} \\ . & . & . & . \\ x_{1p} & x_{2p} & . & x_{np} \end{bmatrix}.\begin{bmatrix} x_{11} & x_{12} & . & x_{1p} \\ x_{21} & x_{22} & . & x_{2p} \\ . & . & . & . \\ x_{n1} & x_{n2} & . & x_{np} \end{bmatrix} = \frac{X`.X}{N}$$

Siendo la matriz X la matriz de datos centrados de individuos por variables, es decir, donde a cada valor se ha restado el valor medio de la variable.

2.2.4.- PROYECCION DE LA NUBE DE INDIVIDUOS Y VARIABLES.

Desde una perspectiva general y teniendo en cuenta la definición de inercia expuesta en el apartado anterior podemos obtener la máxima inercia proyectando un punto sobre un eje de manera que:

Gráfico 2.2.7.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Gráfico de proyección de la nube



Autor: José Aguayo E. Fuente: ACP por Salvador C. Arroyo de la Universidad de Valencia.

La proyección del vector que representa la variable sobre el eje cuyo vector unitario U_1. Aplicando el Teorema de Pitágoras la distancia entre $OA^2 = AB^2 +$

OB^2 luego como queremos minimizar las distancias entre z_i y x_i se tratara de minimizar:

$$\min \sum_{i=1}^{n} \|x_i - z_i\|^2 = \min \sum_{i=1}^{n} \|x_i\|^2 - \min \sum_{i=1}^{n} \|z_i\|^2$$

dado que el primer término del segundo miembro es constante, la minimización se consigue maximizando el segundo, es decir, maximizando la suma de las proyecciones al cuadrado.

Siendo la proyección:

$$z_i = x`_i \cdot U_1 = \sum_{j=1}^{p} x_{ij} u_{1j}$$

Esto nos llevaría a elevar al cuadrado las proyecciones, que matricialmente seria:

$$\mathbf{U_1'X'X\,U_1}$$

Por tanto se debe maximizar $\mathbf{U'_1X'XU_1}$ sujeto a la restricción de que $\mathbf{U'_1U_1 = 1}$

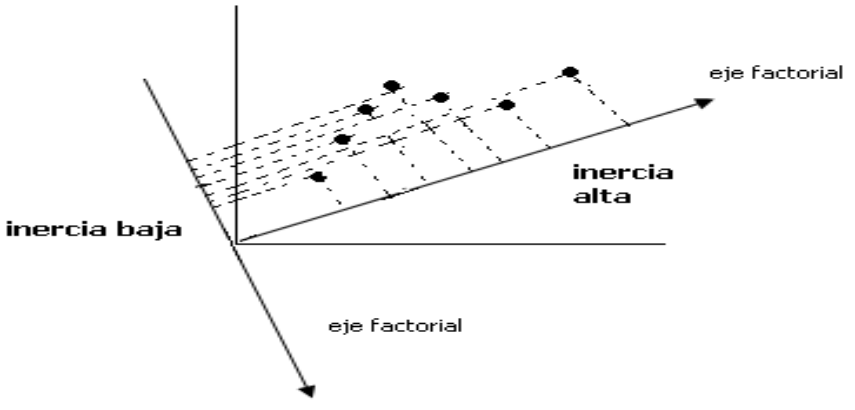
Desde otro punto de vista se puede estudiar la obtención de la máxima inercia a través de la matriz de varianzas-covarianzas.

Sea la tabla de datos n.p. matriz X de individuos por variables y su

representación en forma de nube de puntos-individuo.

Supongamos que conocemos P=2 variables: x_1, x_2. Nuestro objetivo es condensar esa información en una sola variable sintética (factor, componente principal) función de x_1, x_2 que nos represente adecuadamente las dos variables. Es decir interesa reducir la nube de puntos de manera que se obtenga una representación a la vez accesible a nuestra visión y fiel, en el sentido que en la representación de la nube se mantenga el máximo de información que ella contiene. La representación será accesible si se proyecta la nube sobre un subespacio de pequeña dimensión y será fiel si la dispersión de la nube proyectada es casi igual a la nube propiamente dicha.

Gráfico 2.2.8.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Gráfico de la nube proyectada



Autor: José Aguayo E. Fuente: ACP por Salvador C. Arroyo de la Universidad de Valencia**.**

La inercia de la nube proyectada depende del eje que se ha elegido para proyectar y de la inercia que contienen los datos originales.

En general se trata de buscar un subespacio de dimensión m<P y m'<N y esto lleva a encontrar un sistema de vectores U ($u_1 \ldots u_m$**)** y V (($v_1 \ldots v_m$**)** ortonormado para la métrica $R^P R^N$ que engendran el subespacio de manera que sea máxima la inercia de las nubes sobre los subespacios.

La proyección según vimos del vector x_i sobre el eje definido por el vector unitario U es la coordenada z_i del punto en dicho eje:

$$Z_i = X_I.U = \left(x_a x_a \cdot x_{ip}\right)\begin{bmatrix} u_1 \\ u_2 \\ . \\ u_p \end{bmatrix} = x_a u_1 + x_a u_2 + ...x_{ip} u_p = \sum_{j=1}^{p} u_j x_{ij}$$

Cuando hay p variables las coordenadas de las proyecciones de los n individuos sobre el nuevo eje forman un vector columna Z de n elementos: $Z \equiv (z_1, z_2 \ldots z_n)'$ que se calcula multiplicando la matriz X de datos originales por el vector U:

$$Z = XU = \begin{bmatrix} x_{11} & x_{12} & . & x_{1p} \\ x_{21} & . & . & x_{2p} \\ . & . & . & . \\ x_{n1} & x_{n2} & . & x_{np} \end{bmatrix}.\begin{bmatrix} u_1 \\ u_2 \\ . \\ u_p \end{bmatrix}$$

la media de las proyecciones $\bar{z}$ coincide con la proyección del centro de n gravedad de la nube $\bar{x}$.

$$\bar{z} = \frac{1}{N}\sum_{i=1}^{n} z_i = \frac{1}{N}\sum_{i=1}^{n} X_i U = \frac{1}{N}\left[\sum_{i=1}^{n} x_a \sum_{i=1}^{n} x_a \cdot \sum_{i=1}^{n} x_{ip}\right]\begin{bmatrix} u_1 \\ u_2 \\ . \\ u_p \end{bmatrix} =$$

$$= \left[\bar{x}_1 \bar{x}_2 \cdot \bar{y}_p\right]\begin{bmatrix} u_1 \\ u_2 \\ . \\ u_p \end{bmatrix} = \bar{x}u$$

Como la inercia de una nube es su varianza:

INERCIA $(Z)=\sigma_I^2$

$$\sigma_i^2=\frac{1}{N}\sum_{i=1}^{n}\left(z_i-\bar{z}\right)^2=\frac{1}{N}\sum_{i=1}^{n}\left(x_j u-\bar{x}u\right)^2=\frac{1}{N}\sum_{i=1}^{n}\left[\left(x_i-\bar{x}\right)u\right]^2=$$

$$=\frac{1}{N}\sum_{i=1}^{n}\left[U'\left(x_i-\bar{x}\right)\left(x_i-\bar{x}\right)U\right]=\frac{1}{N}U'\left[\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(x_i-\bar{x}\right)\right]U$$

como: $\frac{1}{N}\sum_{j=1}^{n}\left(x_J-\bar{x}\right)'\left(x_J-\bar{x}\right)=S$

ya que según vimos: $S=\frac{X'X}{N}$

La inercia de Z= $\sigma_z^{\ 2}=U'SU$

Hasta ahora lo que hemos realizado es proyectar un conjunto de n individuos sobre un eje. El paso siguiente consiste en proyectar sobre m ejes (m<p), definidos por los m vectores unitarios U_1, U_2…U_m ortogonales. Las proyecciones conforman una matriz **Z** de n filas (una para cada individuo) y m columnas (una para cada eje factorial), y se obtiene multiplicando la matriz de datos **X** por la matriz cuyas columnas son las m vectores unitarios:

$$Z_\alpha=XU_\alpha$$

2.2.4.a.- Obtención del eje de máxima inercia

El primer eje (componente) es la variable sintética o combinación de variables originales que resume mejor la información que contienen.

Se trata de obtener la primera componente principal, es decir, el vector unitario U_1 de manera que U'i Ui =1 que haga máxima la inercia de la nube de puntos proyectada sobre el eje definido por dicho vector (zi).

luego hay que:

Maximizar: $U`_1 S U_1$

Sujeto a que: $U`_1 U_1 = 1$

para obtener este eje el instrumento matemático para maximizar es la diagonalización de una matriz. La función a maximizar sujeta a las restricciones anteriores es:

$$L = U`_1 V U_1 - \lambda_1 \left( U`_1 U - 1 \right)$$

derivando respecto al vector U:

$$\frac{\delta L}{\delta U_1} = 2SU_1 - 2\lambda_1 U_1 = 0$$

y por tanto:

$$SU_1 - \lambda_1 U_1 = 0 \Rightarrow SU_1 = \lambda_1 U_1$$

λ_1 se llama valor propio y nos indica el numero de veces que se amplifica la

longitud del vector U_1 (vector propio).

Los valores propios se obtienen resolviendo la ecuación característica, ecuación polinómica de grado n en general, que se obtiene igualando a cero el determinante de la matriz.

$$\left|s - \lambda_I\right| = 0$$

De la resolución de la ecuación característica, obtenemos los valores propios (tantos como variables originales) de manera que la traza de la matriz diagonal que contiene los valores propios nos indica la máxima inercia que coincide con la varianza total de la nube de puntos.

Para obtener los vectores propios (que nos indican la dirección del eje factorial) asociados a los valores propios, se sustituye el valor propio λ_1 en:

$$\left[s - \lambda_1 1\right] u_1 = 0$$

Sujeto a la restricción $U'_1 U_1 = 1$

resolviendo el sistema se obtiene:

$$U_1 = \begin{bmatrix} u_{11} \\ u_{12} \\ . \\ . \\ . \\ u_{ip} \end{bmatrix}$$

Cada individuo tiene una proyección sobre ese nuevo eje. El conjunto de esas proyecciones sobre el eje obtenido a través del vector U es una nueva variable de manera que las coordenadas del individuo X sobre el eje U_i se obtienen a partir del producto escalar, de manera que obtenemos:

$$Z_a = X_1 U_1 = \sum_{J=1}^{P} x_{i1} u_{11} + ... + x_{ip} u_{ip}$$

Una vez obtenida la primera componente principal, buscamos la segunda. Es decir, el eje definido por el vector U_2 ortogonal con U, que maximice la inercia de la nube proyectada sobre él no condensada por la primera componente. Luego hay que:

Maximizar: $u`_2 s u_2$

Sujeto a: $U`_2 U_2 - 1$

$U`_2 U_1 = 0$

La segunda componente será el vector propio de la matriz S asociado al X_2 (segundo valor propio mayor). A su vez este valor propio refleja la inercia de la proyección.

La proyección de un individuo sobre este eje se representa mediante:

$$Z_a = X_1 U_2 = \sum_{j=1}^{p} x_{1j} u_{2j} = x_a u_{21} + ... + x_{ip} u_{2p}$$

Estos valores constituyen una nueva variable artificial, combinación lineal de las p variables originales. Donde Y_{i2} nos indica las coordenadas del individuo i en el segundo eje factorial.

De forma similar se puede demostrar que los demás componentes son los vectores propios de S asociados a los valores propios ordenados en sentido decreciente. En virtud de las propiedades de la diagonalización de matrices simétricas, si el rango de la matriz S es p habrá p componentes o factores asociados a los p valores propios. Donde cada uno de los valores propios proporciona la parte de la inercia (varianza total) de la nube acaparada por la componente.

En general podemos decir que:

$$Z_{ia} = u_{a1} x_{1i} + u_{a2} x_{2i} + ... + u_{ap} x_{pi}$$

$$Z_{ia} = \sum_{j=1}^{p} x_{ij} u_{aj}$$

Como las componentes son variables centradas se tipifican fácilmente si se divide por su desviación típica. Designaremos Y_α la componente α -esima tipificada definida por el cociente entre Z_α (componente) y la varianza de $Z(\lambda_\alpha)$ de forma que:

$$y_a = \frac{z_a}{\sqrt{\lambda_a}}$$

Luego dividiendo la expresión anterior por la desviación de z obtenemos la expresión

$$Y_{ia} = \frac{z_{ia}}{\sqrt{\lambda_a}} = \frac{u_{a1}}{\sqrt{\lambda_a}} x_{1i} + ... + \frac{u_{ap}}{\sqrt{\lambda_a}} x_{pi}$$

si designamos a $c_{\alpha j} = \frac{u_{\alpha j}}{\sqrt{\lambda_\alpha}}$ obtenemos que:

$$Y_{ia} = C_{a1} X_{1i} + C_{a2} X_{2i} + ... + C_{ap} X_{pi}$$

A la matriz formada por los coeficientes e se le denomina Factor Score. Matriz de puntuaciones factoriales, que se puede expresar como:

$$P = D^{1/2}.U^{`}$$

2.2.4.b.- Estudio a través de la matriz de correlaciones.

Generalmente los paquetes informáticos utilizan la matriz de correlaciones para realizar el ACP. La métrica que se utiliza para poder trabajar con la matriz de correlaciones R e diferencia de la euclidea al introducir el cociente 1/N en la transformación de las variables. De manera que la variable original se transforma de la siguiente manera:

$$x_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{N}\sigma_j}$$

La matriz de los datos originales X se transformara por tanto en otra C de manera que:

C'C= R

En este caso la distancia entre de individuos i i' seria:

$$d^2(i,f)=\frac{1}{N}\sum_{j=1}^{p}\left(\frac{x_{ij}-x_{fj}}{\sigma_f}\right)^2$$

Esto hace coincidir la distancia en esta métrica con la definición dada de inercia.

Por tanto el proceso a seguir será el mismo que hemos realizado cuando trabajábamos con la matriz 5 de varianzas -covarianzas. Aplicando la maximización de la proyección de la nube de puntos descrita anteriormente, habrá que:

maximizar $U_1'RU_1$

con la restricción $U_1' U_1 =1$

Y a partir de ahí obtener las componentes.

Proyección de la nube de puntos-variables en R^N

Análogamente, en el espacio R^N (variables) se tratara de buscar los ejes que maximizan la suma de las proyecciones de las variables al cuadrado.

Sea v_i el vector director del subespacio de dimensión 1 que pasa por el

origen. La proyección de un punto j sobre el eje viene dado por:

$$W_a(j) = X'_j V_1 = \sum x_{ij} v_{ij} = \left[x_{1j} .. x_{nj}\right] \begin{bmatrix} v_{11} \\ \cdot \\ \cdot \\ v_{n1} \end{bmatrix}$$

Los p valores de estas proyecciones son las p fila s del vector de los productos escalares **X'V$_1$** , y por lo tanto la suma de los cuadrados es ahora **V'$_1$XX'V$_1$**. Maximizando esta expresión sometida a la restricción **V'V =1** antes comentada obtendremos (a través de la diagonalización, los valores y vectores propios) de manera que al igual que hacíamos en R^p obtendríamos un subespacio R^m tal que m'<n donde los vectores propios $v_1 \ldots v_n$ asociados a los valores propios $\mu_1 \ldots \mu_n$, están ordenados de mayor a menor.

2.2.5.-RELACION ENTRE LOS DOS ESPACIOS R^p Y R^N (DUALIDAD)

Vamos a ver cual es la relación existente entre estos dos vectores propios (U y V), de tal manera que teniendo uno podamos obtener el otro. De esta manera podremos obtener el subespacio de R^N sin tener que realizar todo el proceso. Además de obtener las funciones que relacionan los dos espacios

Por definición de V_α (vector propio)

$$XX'V_\alpha = \mu_\alpha V_\alpha \text{ z}$$

$$V_\alpha 1/\sqrt{\lambda_\alpha}\ XU_\alpha$$

La proyección de un punto sobre el eje factorial a la representaremos por:

$$W_\alpha(j) = X\,\text{`}V_\alpha = X\,\text{`}\frac{1}{\sqrt{\gamma_\alpha}}XU_\alpha =$$

$$= \frac{1}{\sqrt{\lambda_\alpha}}X\,\text{`}XU_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\lambda_\alpha U_\alpha = \sqrt{\lambda_\alpha}U_\alpha$$

Podemos establecer por tanto las relaciones de transición que relacionan los dos espacios de manera que, si imponemos que:

u' u_α =v'$_\alpha$ =1 se obtiene según vimos que k_α= k'$_\alpha$=1/$\sqrt{\lambda_\alpha}$.

Podemos pues ahora escribir el siguiente sistema de relaciones fundamentales

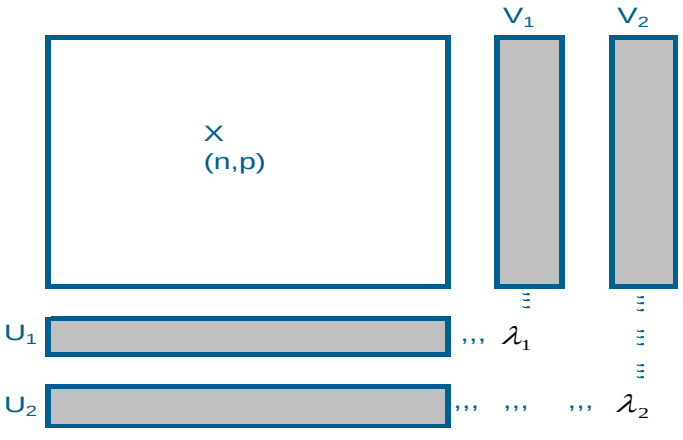
$$U_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}X\,\text{`}V_\alpha$$

$$V_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}XU_\alpha$$

Gráfico 2.2.9.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

RELACION ENTRE LOS DOS ESPACIOS R^p Y R^N



Autor: José Aguayo E. Fuente: ACP por Salvador C. Arroyo de la Universidad de Valencia.

El eje Z_α soporte del vector unitario U_α, se llama el α-ésimo eje factorial de R^p. El W_α soporte del vector unitario V_α es el α-ésimo eje factorial de R^N.

Las coordenadas de los puntos de la nube sobre el eje α en R^p son, por construcción, las coordenadas de XU_α , y las de los puntos de la nube sobre el eje α en R^N son, por construcción, las coordenadas de $X'V_\alpha$

Por tanto las relaciones son

$$W_\alpha(j) = X`V_\alpha = X`\frac{1}{\sqrt{\lambda_\alpha}} XU_\alpha = X`\frac{1}{\sqrt{\lambda_\alpha}} Z_\alpha(i) = \frac{1}{\sqrt{\lambda_\alpha}} X`Z_\alpha(i)$$

$$Z_\alpha(i) = XU_\alpha = X\frac{1}{\sqrt{\lambda_\alpha} X`V_\alpha} = X\frac{1}{\sqrt{\lambda_\varepsilon} W_\alpha}(j) = \frac{1}{\sqrt{\lambda_\alpha}} XW_\alpha(j)$$

$$W_\alpha(j) = \frac{1}{\sqrt{\lambda_\alpha}} X`Z_\alpha(i)$$

$$Z_\alpha(i) = \frac{1}{\sqrt{\lambda_\alpha}} XW_\alpha(j)$$

También a partir de las relaciones:

$$U_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X`V_\alpha$$

$$V_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} XU_\alpha$$

Se puede reconstruir la matriz original:

$$\sqrt{\lambda_\alpha}U_\alpha = X`V_\alpha$$
$$\sqrt{\lambda_\alpha}V_\alpha = XU_\alpha$$

si postmultiplicamos los dos miembros por V'$_\alpha$ y U '$_\alpha$ respectivamente tenemos:

$$\sqrt{\lambda_\alpha}U_\alpha V`_\alpha = X`V_\alpha V`_\alpha$$
$$\sqrt{\lambda_\alpha}V_\alpha U`_\alpha = XU_\alpha U`_\alpha$$

y sumemos para todos los valores α (si existen valotes propios nulos, los vectores U_α correspondientes completan la base de R^p). (igualmente los vectores V_α completan la base R^N).

$$X = \sqrt{\lambda_\alpha}V_\alpha U`_\alpha$$
$$X`(j) = \sum_{\alpha=1}^{p}\sqrt{\lambda_\alpha}u_{\alpha j}jv`_{\alpha j}$$

2.2.6.- CORRELACION ENTRE FACTORES Y VARIABLES ORIGINALES

Una de las razones para realizar una ACP es sintetizar un conjunto de variables en otras de inferior dimensión. En definitiva tenemos que obtener esas nuevas variables y darles un nombre, para hacerlo debemos interpretar el significado en el contexto del problema que se analiza y debemos conocer cual es la correlación entre estas nuevas variables sintéticas y las variables originales.

En general cada factor estará muy correlacionado con alguna o algunas variables

y menos con las demás, por ello es importante saber las correlaciones entre los factores y las variables originales.

El coeficiente de correlación lineal de Pearson es el cociente entre la covarianza y el producto de las desviaciones típicas.

La covarianza maestral entre X_j y Zα viene dada por la expresión:

$$Cov(Z_a, X_j) = \frac{1}{N}\sum_{i=1}^{n} x_{ij} z_{ia} \quad \begin{matrix}(i = 1.....n)\\(a = 1.....p)\end{matrix}$$

Donde x_{ij} ; y $z_{iα}$ son valores centrados. La expresión anterior se puede escribir en forma matricial como:

$$Cov(Z_a, X_j) = \frac{1}{N} X`_J Z_a = \frac{1}{N} X`_j \cdot Z_a = \frac{1}{N} Z`_a \cdot X_j \quad \begin{matrix}\forall j = 1...., p\\ \forall a = 1..., F\end{matrix}$$

donde el vector $\mathbf{x_j}$ es columna de la tabla de datos centrados, y por lo tanto X_j es la fila j-ésima de la matriz transpuesta (X'). El vector $\mathbf{x_j}$ se puede expresar en función de la matriz X, utilizando el vector de orden p, al que le designaremos por δ, que definiremos como un vector fila con todos sus elementos nulos excepto el j-ésimo que vale 1, así podemos escribir:

$$X`_j = \delta X` = \begin{bmatrix}0 & . & 1 & . & 0\end{bmatrix} \begin{bmatrix} x_{11} & . & . & x_{il} & . & x_{nl} \\ . & . & . & . & . & . \\ x_{1j} & . & . & x_{if} & . & x_{nj} \\ . & . & . & . & . & . \\ x_{1p} & . & . & x_{ip} & . & x_{np} \end{bmatrix} = \begin{bmatrix} x_{1j} & x_{2j} & .. & .. & x_{nj} \end{bmatrix}$$

Teniendo en cuenta que Z_α=X U_α la covarianza se puede expresar de la siguiente manera

$$Cov(Z_a, X_j) = \frac{1}{N}\delta`X`XU_a$$

Teniendo en cuenta que (X'X)/N = S y que U_α es el α-esimo vector propio de S, la última expresión se convierte en:

$$\frac{1}{N}\delta`X`XU_a = \delta`SU_a = \lambda_a(0 \quad . \quad 1 \quad . \quad 0)U_a = \lambda_a u_{aj}$$

Esta expresión indica que la covarianza entre la variable j y el factor α es directamente proporcional a la inercia del factor y al j-ésimo elemento del vector unitario que lo define. Por tanto la correlación entre ambas será:

$$corr(a, j) = \frac{\text{cov}(a, j)}{\sigma_a \sigma_j} = \frac{\lambda_a u_{aj}}{\sigma_j \sqrt{\lambda_a}} = \frac{\sqrt{\lambda_a} u_{aj}}{\sigma_j}$$

La correlación componente - variable depende de la desviación típica (en caso de que las variables originales no estén tipificadas), y por consiguiente depende de la unidad de medida que se utilice: la interpretación de las componentes varía si cambiamos de medida las variables.

Si utilizamos variables tipificadas su desviación típica de la variable será uno y por tanto:

$$corr(a.j) = \sqrt{\lambda_a} u_{aj}$$

Estos serán los valores de las coordenadas de las variables en los ejes factoriales, en el supuesto de un análisis de componentes principales normado (análisis a partir de la matriz de correlaciones R).

A la matriz formada por las coordenadas factoriales de las variables (proyecciones de las variables sobre los ejes factoriales se le denomina FACTOR MATRIX).

Por otro lado tanto la traza de la matriz R (de correlaciones) como la de la matriz S de varianzas -covarianzas es igual a la suma de la diagonal principal de la matriz de valores propios.

$$tr(R) = tr(S) = \sum_{j=1}^{p} r_{ij} = \sum_{j=1}^{p} s \sum ij = \sum_{j=1}^{p} \lambda_a$$

Siendo la traza de S la varianza total del sistema.

Hay que entender que la matriz de valores propios tiene toda la información relevante de la matriz de varianzas-covarianzas S y la matriz de correlaciones R en la diagonal, siendo redundante la información que aportan los elementos exteriores de la diagonal.

Cada valor propio se interpreta no solo como la varianza de la componente, sino

también como la parte de la varianza que el α-esimo eje principal explica y el ratio:

$$\frac{\lambda_a}{p} = \frac{\lambda_a}{tr(S)}$$

como índice de la importancia de esta componente en la varianza total en el conjunto de las variables originales. De esta manera podemos obtener el porcentaje de la inercia total de la nube explicada por cada una de las componentes.

2.2.7.- NUMERO DE COMPONENTES A RETENER

La matriz factorial puede presentar un número de factores superior al necesario para explicar la estructura de los datos originales. Generalmente hay un conjunto reducido de factores, los primeros, que son los que explican la mayor parte de la variabilidad total. Los otros factores suelen contribuir relativamente poco. Uno de los problemas que se plantean, por tanto, consiste en determinar el número de componentes que debemos conservar.

Si retuviésemos todos las componentes (que seria igual al número de variables originales), entonces la matriz **Z** de coordenadas de la n individuos en los p componentes contendría toda la información (inercia) de la matriz X de datos

originales.

Sin embargo nuestro objetivo es reducir la dimensión del problema. Por tanto debemos retener los m primeros factores (m<p), excluyendo los restantes.

El porcentaje de inercia condensado por estos m factores conjuntamente es:

$$\frac{\sum_{a=1}^{m} \lambda_a}{tr(S)} \cdot 100$$

$$\frac{\sum_{a=1}^{m} \lambda_a}{p} .100$$

Si las variables originales no estuvieran correlacionadas, o bien si la correlación de todos los pares de variables fuera idéntica, entonces cada valor propio seria igual a la unidad, de forma que cada factor explicarla exactamente la misma cantidad de inercia que cada variable. En este caso seria imposible reducir el numero p de variables y encontrar por tanto un subespacio menor de p.

Generalmente las variables originales presentan correlaciones, de forma que los valores propios aparecen ordenados en sentido decreciente. Las componentes explican entonces distintos porcentajes de la varianza total (inercia), de forma que reteniendo los m primeros se consigue dos objetivos:

- Se conserva gran parte de la inercia total de la nube de puntos

Conociendo la matriz **Z** (nxm) de coordenadas de los individuos en los factores retenidos, se puede reproducir de forma aproximada la matriz de correlación original.

La decisión de cuantos factores deben retenerse depende del tipo de problema que estemos analizando, de la precisión requerida, de la interpretabilidad de los componentes, etc... . Se trata de explicar la máxima varianza de la nube de puntos (variables originales) con el mínimo de factores.

Uno de los criterios más conocidos y utilizados es el criterio o regla de Kaiser, que indicaría “hay que conservar solamente aquellos factores cuyos valores propios son mayores a la unidad”. Este criterio es el que suelen utilizar los programas estadísticos por defecto. Sin embargo este criterio tiende a sobrestimar el Nº de factores.

Otro criterio es el Scree-Test de Cattell, consiste en representar en un sistema de ejes los valores propios (ordenadas) y el número de factores (abscisas). Sobre la gráfica resultante se traza una línea recta base a la altura de los últimos valores propios (más pequeños) y aquellos que queden por encima indicaran el n° de

factores a retener.

Es utilizado para estudiar el decrecimiento de los valores propios. El principio de lectura de este gráfico es el siguiente: si dos factores están asociados a valores propios casi iguales, representan la misma parte de variabilidad y no hay lugar, a priori, para retener uno y no el otro en la interpretación. Recíprocamente, un fuerte decrecimiento entre dos valores propios sucesivos incita a retener en la interpretación los factores precedentes a este decrecimiento.

En la practica, se observa a menudo el fenómeno siguiente: los f primeros valores propios presentan un decrecimiento bastante irregular; después, mas allá del rango f, el decrecimiento es muy regular. Esto indica que los f primeros factores corresponden cada uno de ellos a irregularidades en la forma de la nube de puntos estudiada que requieren ser interpretadas y sugiere que los factores siguientes no representan más que el inevitable ruido que acompaña a toda observación de naturaleza estadística.

Caso extremo, un decrecimiento regular desde el primer valor propio traduce una nube casi esférica y, por tanto, datos poco estructurados de los que los factores son poco sintéticos. Un diagrama de este tipo presagia un interés limitado de los

factores.

Velieer, propone el método MAP (Minimum Average Partial), que implica calcular el promedio de las correlaciones parciales al cuadrado después de que cada una de las componentes ha sido parcializado de las variables originales. Cuando el promedio de las correlaciones parciales al cuadrado alcanza un mínimo no se extraen más componentes. Este mínimo se alcanza cuando la matriz residual se acerca más a una matriz identidad. Un requisito para utilizar esta regla es que cada una de las componentes retenidas debe tener al menos dos variables con pesos altos en ellos.

El análisis paralelo fue sugerido por Horn, quien señala que a nivel poblacional los autovalores de una matriz de correlaciones para variables no correlacionadas tomarían valor 1. Cuando se generan matrices maestrales basadas en esa matriz poblacional por fluctuaciones debidas al azar los autovalores excederán levemente de 1 y los últimos estarán ligeramente por debajo de 1. Hora propone contrastar los autovalores encontrados empíricamente en los datos reales con los obtenidos a partir de una matriz de variables no correlacionadas basadas en el mismo número de variables que los datos empíricos y en el mismo tamaño de la muestra. Los componentes empíricos con autovalores superiores a los de la

matriz son retenidos.

En general se deben retener, sin perder de vista los criterios objetivos, aquellos factores que se saben interpretar. Seria perjudicial rechazar, con criterios estadísticos, un factor que se sabe interpretar y seria delicado retener un factor que no se sabe interpretar.

2.2.8.- ELEMENTOS SUPLEMENTARIOS.

El objetivo del análisis es reducir el espacio conformado por los individuos y variables originales en un subespacio de dimensión menor manteniendo la máxima información sin modificar la dispersión inicial.

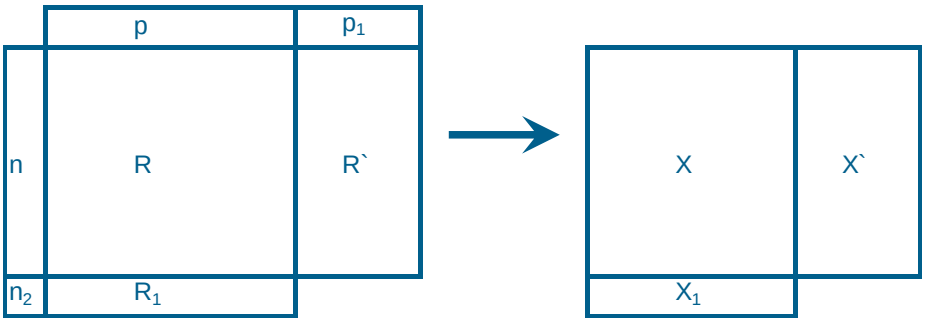
A esas variables e individuos se les denominan elementos activos. Sin embargo también pueden posicionarse en ese mismo sub -espacio, otros elementos (puntos-fila o puntos-columna de la matriz de datos) que no han participado en la construcción de los ejes factoriales y que son llamados elementos suplementarios o ilustrativos.

Los elementos suplementarios intervienen a posterior para caracterizar a los ejes.

Su introducción en el análisis constituye una aportación fundamental que permitirá enriquecer la interpretación de los factores.

La tabla de datos **R** puede ser así completada en columna por una tabla de n líneas y con p_s columnas $\mathbf{R^+}$ y en línea en una tabla $\mathbf{R_+}$ con n_s líneas p_s columnas.

Gráfico 2.2.10.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Gráfico elementos suplementarios



Autor: José Aguayo E. Fuente: ACP por Salvador C. Arroyo de la Universidad de Valencia.

Las tablas $\mathbf{R^+}$ y $\mathbf{R_+}$ pueden transformarse respectivamente en tablas $\mathbf{X^+}$ y $\mathbf{X_+}$.

Se llama individuo suplementario o ilustrativo a aquellos individuos que no deseando que intervengan en la determinación de los ejes interesa conocer la

posición de su proyección sobre los ejes obtenidos por el resto de la población

Para situar los individuos suplementarios en el espacio R^p es necesario realizar la transformación:

$$x_{+ij} = \frac{x_{+ij} - \overline{x_j}}{\sqrt{N}.\sigma_j}$$

Las coordenadas de los nuevos puntos individuos son:

$$Coord(i,a) = x_{+j} \cdot u$$

De igual manera las variables suplementarias o ilustrativas son aquellas que se proyectan sobre los ejes determinados por las restantes variables.

En R^N para que las distancias entre variables se interpreten en término de correlación las variables deben ser continuas y es indispensable efectuar la transformación:

$$x_{ij}^{+} = \frac{x_{ij}^{+} - \overline{x}_j^{+}}{\sqrt{N} \cdot \sigma_j^{+}}$$

Donde se calcula la nueva media y desviación incorporando las nuevas variables suplementarias a las variables activas; de esta forma se puede posicionar esta variable ilustrativa sobre la esfera de radio 1.

$$Coord(j,a) = x^{+} \cdot v_a$$

Las relaciones de transición nos permiten determinar los vectores propios $\mathbf{y_0}$ a partir de los valores propios λ_α y las coordenadas factoriales de los individuos activos sin que sea necesario realizar las operaciones de las diagonalización, así:

$$v_a = \frac{1}{\sqrt{\lambda_a}} \cdot coord(i,a)$$

Si existe una fuerte correlación entre una componente principal y una variable suplementaria, esta variable caracteriza la componente de una manera más fuerte que las variables activas, ya que las variables suplementarias no han formado parte en la formación de la componente principal.

2.2.9.- CONTRIBUCIONES ABSOLUTAS Y RELATIVAS: ayudas a la interpretación.-

Hasta ahora sabemos cual es la contribución de un eje factorial a la varianza total de la nube de puntos. También que un factor se interpreta a partir de su correlación con las variables originales y que esta correlación es precisamente la proyección de la variable sobre el factor, es decir, las coordenadas de las variables sobre el factor lo que constituye la matriz de saturaciones.

Es indispensable utilizar las ayudas a la interpretación. Un análisis factorial explicado solamente sobre el examen de las gráficas obtenidas tiene un fuerte riesgo de ser erróneo.

Para interpretar un eje es importante saber cuales son los puntos que mas contribuyen a la formación de los ejes. Estas contribuciones son las que se denominan como absolutas y relativas.

Contribuciones absolutas de las variables

Llamaremos contribución absoluta a la aportación de cada una de las variables a la inercia o varianza de cada eje factorial. Es decir nos indicara la contribución o porcentaje de una variable original a la construcción del factor que previamente nos ha indicado a partir de su valor propio el porcentaje de la varianza total de la nube de puntos.

Expresaremos la contribución absoluta como:

$$CAb_a(j) = \frac{Z_a^2(j)}{\sum_{j=1}^{p} Z_a^2(j)} = \frac{coord^2(j,a)}{\sum coord^2(j,a)}$$

como el peso de las variables es 1 para todas ellas, tenemos que:

$$\sum m_j \cdot coord^2(j,a) = \lambda_a$$

$$CAb_a 8 j) = \frac{coord^2(j,a)}{\lambda_a}$$

Podemos comprobar que:

$$\sum_{j=1}^{p} CAb_a(j) = 1$$

También podríamos expresar la contribución absoluta teniendo en cuenta que:

$$Z_a(j) = \sqrt{\lambda_a} u_{aj} = r_{aj}.$$

Como determinamos que los coeficientes de correlación entre variable y factor en ACP coincidían con las saturaciones en AF, tenemos:

$$CAb_a(j) = \frac{r_{aj}^2}{\sum_{j=1}^{p} r_{aj}^2} = \frac{a_{aj}^2}{\sum_{j=1}^{p} a_{aj}^2} = \frac{a_{aj}^2}{\lambda_a}$$

Contribución Relativa de las variables

Expresan la contribución de un factor a la explicación de la dispersión de una variable. No debe confundirse este concepto y creer que es un porcentaje de la contribución absoluta porque estas últimas se pueden medir en tantos por 1 o en porcentaje. Mientras que las contribuciones absolutas permiten saber que variables son las responsables de la construcción de un eje factorial, las contribuciones relativas muestran cuales son las características exclusivas de ese factor.

Matemáticamente hablando los ejes factoriales constituyen bases ortonormales. El cuadrado de la distancia de un punto al centro de gravedad se descompone en suma de cuadrados de las coordenadas en estos ejes. Para un punto, se tiene que:

$$C\,\mathrm{Re}_a(j) = \frac{Z_a^2(j)}{d^2(j,G)} = \frac{coord^2(j,a)}{d^2(j,O)}$$

Téngase en cuenta que la cantidad de información se mide por la suma de las distancias al origen al cuadrado y que al estar la nube centrada sobre el origen la distancia en la métrica empleada (no euclidea) nos da siempre la unidad (la nube de puntos esta repartida sobre una esfera de radio la unidad). Por tanto así:

$$C\,\mathrm{Re}_a(j) = \frac{Z_a^2(j)}{1} = coord^2(j,a)$$

luego las contribuciones relativas coinciden con las coordenadas de cada variable. Al igual que las variables en el caso de individuos tendríamos:

$$CAb_a(i) = \frac{m_i \cdot coord^2(i,a)}{\lambda_a}$$

nos indicara la contribución de cada individuo a la formación del eje factorial.

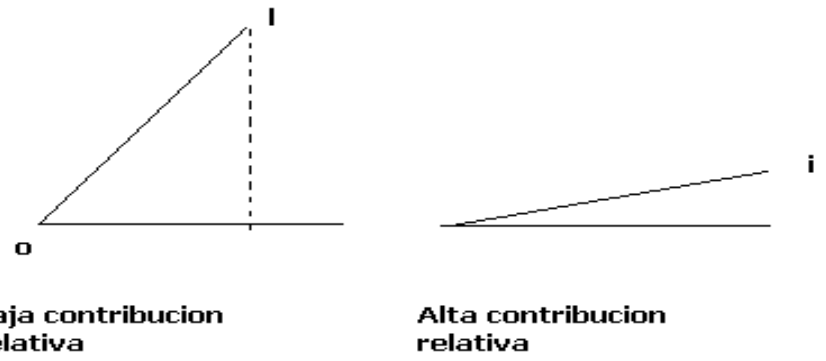
La contribución relativa en el caso de individuos permite evaluar la calidad de la representación de los individuos sobre el eje factorial.

$$C\,\mathrm{Re}_a(i) = \frac{coord^2(i,a)}{d^2(i,G)}$$

A fin de obtener la contribución relativa se calcula la distancia de cada punto individuo al origen. Esta distancia es:

$$d^2(i,G) = \sum_{j=1}^{p} x_{ij}^2$$

Gráfico 2.2.11.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Gráfico de contribución relativa



Autor: José Aguayo E. Fuente: ACP por Salvador C. Arroyo de la Universidad de Valencia**.**

La suma de todas las contribuciones relativas (de cada factor) de un individuo será la unidad. Si lo expresamos en porcentaje el 100%.

$$\sum_{a=1}^{p} C\,\mathrm{Re}_a(i) = 1$$

2.2.10.- INTERPRETACION DE LOS FACTORES.

Los factores se escogen en el orden decreciente de los valores propios. Pueden ser estudiados separadamente o dos a dos con la ayuda de los planos factoriales. Es preciso tener en cuenta que el factor de orden f(f>1) traduce las tendencias residuales no tenidas en cuenta por los factores precedentes.

La interpretación siempre se realiza de forma personal, es el analista quien de forma particular, en función de sus conocimientos y experiencia interpreta unos resultados que otro haría de diferente forma.

Debido a las relaciones existentes entre los dos espacios R^p y R^N , a menudo es preciso consultar alternativamente los resultados relativos tanto a los individuos como a las variables.

En general es más fácil comenzar por el estudio de las variables, ya que en la mayoría de los casos son inferiores al número de individuos y tienen más sentido que los individuos. De esta manera se da mayor relevancia a los datos que han participado directamente en la construcción de los factores

2.2.10.a.- Interpretación de la nube de variables.

Retomando el objetivo principal del análisis de componentes principales, como la obtención de unas nuevas variables sintéticas (factores) combinación lineal de las variables originales, de manera que, sinteticen la información manteniendo la estructura original. El problema que se nos plantea es dar nombre a esas nuevas variables de manera que indiquen fielmente aquellas variables originales que han contribuido principalmente a su construcción. Para ello, en una primera fase, partiremos de la correlación existente entre los factores y las variables originales.

Cada factor estará muy correlacionado con algunas variables de forma que podremos atribuir un significado a las componentes si conocemos estas correlaciones.

Al interpretar eje por eje se consideran las variables activas mas ligadas a cada eje. De esta manera se pueden presentar dos situaciones:

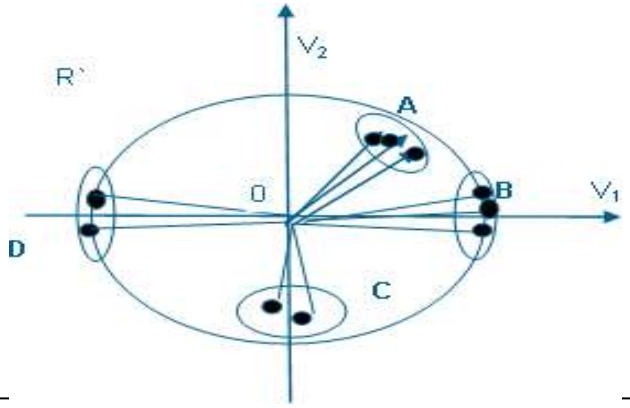
-Todas las variables muy ligadas al factor se sitúen en un mismo lado del eje. El factor aparece entonces como una síntesis entre estas variables.

- Las variables muy ligadas al factor presenten una coordenada positiva para

unas y negativa para las demás. Es preciso entonces buscar un denominador común que, a la vez, relacione las variables situadas en un mismo lado y oponga las variables situadas en diferentes materias, un factor puede traducir la oposición entre materias. Esta fase permite obtener ya la significación general de algunos ejes.

Es interesante ayudar a la interpretación trazando un círculo de radio 1, o círculo de correlaciones, porque la proximidad de un punto al círculo permite juzgar la calidad de las variables. Por otro lado, si unimos los puntos de las variables con el origen visualizamos los ángulos que miden la relación entre las variables bien representadas (próximas al círculo de correlación). Así es posible reagrupar visualmente variables relacionadas entre si y bosquejar de este modo una tipología de las variables.

Gráfico 2.2.12.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Interpretación de nube de variables



Autor: José Aguayo E. Fuente: ACP por Salvador C. Arroyo de la Universidad de Valencia.

La nube de variables esta situada sobre una esfera (0,1), la imagen de los puntos de la nube están situados sobre un plano factorial en el interior de un circulo (0,1) (Gráfico Nº 3).Los puntos de la nube mejor representados para los planos son los que su imagen está más próxima al borde del círculo. Si observamos gráfico, sobre el eje VI el grupo B de variables tiene las coordenadas próximas a 1 y el grupo D próximas a —1; otro grupo se encuentra muy cerca del borde del circulo sin tener sus coordenadas elevadas ni sobre el eje V_1 ni sobre V_2 (grupo A); y por ultimo un grupo con coordenadas relativamente próximas a -1 sobre el eje V_2 (grupo C).

Se dirá que el eje V_1 opone las variables del grupo B a las del grupo D. Cada uno

de estos grupos están formados por variables fuertemente correlacionadas entre si. Se puede interpretar $F_1(i)$ como una nueva variable definida sobre la población considerada y que será función lineal creciente de cada una de las variables del grupo B y función lineal decreciente de cada una de las variables del grupo D. Así el análisis nos aporta dos resultados: los grupos B y D y una nueva variable F_1 (i) que puede sustituir a cada una de estas variables sin que se pierda mucha información.

Como nosotros le hemos dado a cada punto variable un peso igual a la unidad y como la nube de variables esta situada en la esfera de radio 1, el uso de las ayudas a la interpretación no es indispensable, es suficiente con la lectura de las coordenadas de las variables.

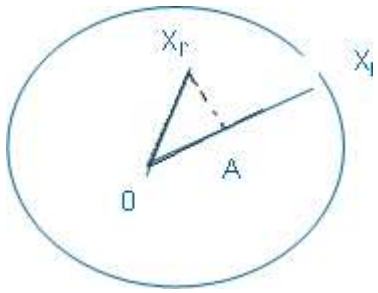
Por otro lado las variables del grupo A están bien representadas para el plano (1,2) sin que estén ,como decíamos, bien representadas para el eje y_1 ni para el eje V_2.

Recordemos que si la variable j esta representada por un punto muy próximo al borde del círculo, se puede ver directamente sobre el gráfico el coeficiente de correlación r_{jj} con otra variable j' cualquiera; es suficiente trazar sobre O_j la

perpendicular resultante de la proyección de x_j así se verifica que:

$$\overline{OA} = x`_i \cdot x_{i`} = r_{ii`}$$

Gráfico 2.2.13.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Correlaciones entre variables



Autor: José Aguayo E. Fuente: ACP por Salvador C. Arroyo de la Universidad de Valencia**.**

El grupo C esta formado por variables que tienen una correlación con las variables de los grupos B y C nula y que tienen una correlación negativa con las variables del grupo A. También podemos decir, que esta próximo al borde del circulo en la dirección negativa del eje V_2. La variable F_2 (i) esta correlacionada negativamente con las variables del grupo C; esta correlación es inferior a la existente entre las variables del grupo B y C con el eje F_1(i).

2.2.10.b.- Interpretación de la nube de individuos

Al contrario de lo que ocurre en la nube de variables los puntos-individuos no está inscrito en un círculo de radio 1. La nube de individuos estará centrada en el origen por la transformación que hemos realizado en los datos iniciales.

A fin de facilitar la interpretación de los resultados del análisis, se calcula a menudo la contribución de cada individuo a la inercia de las proyecciones sobre el eje factorial a. Esta contribución (absoluta) nos indica la contribución de un punto-individuo a la formación del eje factorial. Cuando un individuo posee una CTAb muy alta es aconsejable estudiar con detalle sus características.

Sin embargo, no se puede ver la calidad de la representación de un punto sobre un plano, no depende solo de su distancia al origen, ya que, en el espacio R^p no están todos los puntos a la misma distancia del origen. Se puede obtener una medida de la calidad de la representación y posición del individuo i sobre el eje α, a través de la contribución relativa. Ella nos indica la contribución del eje factorial a la distancia entre el punto-individuo y el origen.

Nos interesan los individuos que tienen las Cre. más altas. Como los individuos

están afectados por un mismo peso la inercia de un punto varia según su distancia al centro de gravedad y los individuos que contribuyen mas serán los más alejados.

La segunda fase de interpretación se realiza mediante los planos factoriales. Comparativamente a la fase precedente, el plano factorial aporta el poder sintético del gráfico más sugestivo que una lista de coordenadas y la consideración simultánea de dos dimensiones que da una imagen fiel de los datos y puede también sugerir la interpretación de otras direcciones además de los ejes factoriales.

La construcción de los planos factoriales pasa por establecer que factores debemos cruzar. Para ello, podemos tomar como referencia:

- La inercia asociada a los factores. Se cruzan preferentemente factores asociados a valores propios más próximos. Por ello se tiende a construir la sucesión de planos que cruzan los factores 1 y2, los factores2 y 3…
- La experiencia y conocimiento del analista sobre las variables y el entorno conceptual de las variables, tanto de las originales como de las nuevas (factores)
- La significación general del factor. Se puede desear poner el foco de atención

en algunas variables y por tanto en los planos en los que estas se encuentran bien representadas.

2.2.10.c.- Representación simultanea de variables e individuos.

La representación gráfica ideal que resume todo el análisis es por excelencia la representación simultánea de individuos y variables. La disposición de las variables permite interpretar la nube de individuos de manera que son las variables las responsables de las proximidades entre los individuos.

No se puede interpretar la distancia entre un individuo y una variable, ya que, sus proyecciones no están medidas en la misma escala. Si se pueden estudiar las posiciones relativas de los individuos respecto de una variable.

En el espacio R^p de los n puntos-individuos una vez realizada la transformación de la tabla de datos, se dispone de dos sistemas de ejes. Los antiguos ejes unitarios δ correspondientes a las p variables antes del análisis y los nuevos ejes unitarios u_0 constituidos por los ejes factoriales. La posibilidad de una representación simultánea reside entonces en la proyección (en línea suplementaria) del antiguo eje sobre el nuevo eje.

Así es posible representar en R^p las direcciones dadas por las variables sobre el plano factorial de la nube de individuos y se interpreta el vector que une el origen con el punto como la dirección de alargamiento. Un individuo bien representado alejado en la dirección de la variable j, indica que ese individuo toma un valor mas alto que la media en esa variable.

Para representar simultáneamente individuos y variables se aplicara a las coordenadas de los individuos un coeficiente corrector (raíz de n/p) permitiendo de esta forma una representación equilibrada de las nubes.

La proyección de la nube de individuos vendar determinada por la ecuación:

$$Z_a(i) = \sqrt{\frac{n}{p}} \cdot X \cdot u_a$$

Donde n es el numero de individuos y p el Nº de variables. La proyección de los puntos variables vendrá determinada por:

$$W_a(j) = X` \cdot v_a = \sqrt{\lambda_a} \cdot u_a$$

2.2.11.- OPTIMIZACION DE LA MATRIZ FACTORIAL A TRAVES DE UNA ROTACION

La matriz factorial indica la relación entre los factores y las variables. Sin embargo, a partir de ella en muchas ocasiones resulta difícil la interpretación de los factores. Mejoraríamos esa interpretación, si obtuviéramos unos factores, de manera, que cada variable o pequeño grupo de variables originales tenga una correlación lo más próxima a 1 que sea posible con uno de los factores y correlaciones próximas a O con el resto de los factores. De esta forma cada factor tendría una correlación alta con un grupo de variables y baja con el resto.

El procedimiento para mejorar la interpretación es a través de una transformación. Esta transformación que se realiza sobre los factores se denomina genéricamente rotación. En síntesis consiste en hacer girar los ejes de coordenadas que representan a los factores hasta conseguir que se aproxime al máximo a las variables en que están saturados.

A través de la rotación se pretende eliminar, por un lado, los signos negativos de la matriz de saturaciones (A) y por otro, obtener que los factores estén muy correlacionados con un grupo de variables y poco con las demás.

En general se trata de obtener una matriz de transformación T, de orden (m,m) tal que:

$$B = A \cdot T$$

Si Atiene inversa, que es el caso del método de factores principales, la matriz T es:

$$T = A^{-1} \cdot B$$

pero normalmente A es de orden (p, m) y no tiene inversa. El problema se resuelve planteando unos criterios sobre la matriz B y a continuación se halla la transformación adecuada.

La saturación de factores transforma la matriz inicial en otra denominada matriz factorial rotada, de más fácil interpretación. La matriz rotada es una combinación lineal de la primera y explica la misma cantidad de varianza inicial. El objetivo de la rotación es obtener una solución más interpretable. Una forma de conseguirlo es intentando aproximarla al principio de estructura simple, según este principio la matriz factorial debe reunir las siguientes características:

- Cada factor debe tener unos pocos pesos altos y los otros próximos a cero.
- Cada variable no debe estar saturada más que en un factor.
- No deben existir factores con la misma distribución.

Estos tres principios, en la práctica no suelen lograrse, se debe obtener la solución mas aproximada posible a ello.

Con la rotación factorial aunque cambie la matriz las comunalidades no se alteran, sin embargo, cambia la varianza explicada por cada factor.

Existen varios métodos de rotación que podemos agrupar en dos grandes tipos: Ortogonales y Oblìcuos.

2.2.11.a.- Rotación Ortogonal:

En este tipo de rotación los ejes factoriales al rotarlos tienen que mantener la perpendicularidad entre ellos. De esta manera persiste la incorrelación entre los factores. Una rotación ortogonal mantiene la calidad global de la representación, es decir, la capacidad del análisis para sintetizar los datos, así como las comunalidades. Sin embargo varían las correlaciones entre factores y variables, y el porcentaje de inercia condensado en cada factor. Los nuevos factores que se obtienen de una rotación han de ser interpretados y nombrados observando sus correlaciones con las variables. Los factores antiguos y los rotados están correlacionados. El grado que alcanza su correlación depende del ángulo de giro.

Para realizar una rotación ortogonal debemos plantearnos que: dada la matriz factorial A, hallar una matriz ortogonal T, de modo que obtendremos una matriz B (siendo B = AT) sea la matriz factorial de unos nuevos factores ortogonales. En este sentido, dado un numero m de factores (m>l), el conjunto de saturaciones en A no es único, pues toda transformación ortogonal de A proporciona matrices equivalentes:

$$B = A \cdot T$$
$$T \cdot T` = 1$$

La equivalencia se manifiesta en que ambas matrices Ay B reproducen igualmente la matriz R de correlación inicial.

Ambas matrices conservan el mismo porcentaje de varianza proyectado sobre el conjunto de los factores, y se mantendría la varianza explicada de cada una de las variables originales.

El método de rotación ortogonal más conocido es el Varimax. Este método desarrollado por Kaiser, simplifica las columnas de la matriz de factores de manera que obtiene unas correlaciones altas entre los ejes rotados y unas pocas variables y correlaciones prácticamente nulas con el resto. Para ello utiliza el criterio que denomina simplicidad de un factor, midiéndola como la varianza de los cuadrados de sus saturaciones en las variables observables.

Se puede calcular la varianza a partir de los momentos respecto al origen. La simplicidad $S^2{}_\alpha$(j) del factor Z_α(j) será pues:

$$S_A^2(J) = \frac{1}{p}\sum_{j=1}^{p}\left(a_{ja}^2\right)^2 - \frac{1}{p^2}\left(\sum_{j=1}^{p} a_{ja}^2\right)^2$$

Kaiser pretende obtener B =AT de modo que la suma de las simplicidades de todos los factores sea máxima:

$$S^2 = \sum_{a=1}^{n} S_a^2 = \max$$

Este criterio planteaba un problema y es que las comunalidades altas dan lugar a saturaciones altas y las comunalidades bajas a saturaciones bajas, distorsionando el efecto de la rotación. Para evitarlo se aplica lo que llamamos proceso de normalización de Kaiser.

Consiste en normalizar las saturaciones de un factor, dividiéndolas por la raíz cuadrada de su comunalidad. Así la simplicidad del factor Z_α(j) será:

$$S_a^2(j) = \frac{1}{p}\sum_{j=1}^{p}\left(\frac{a_{ja}^2}{h_j^2}\right)^2 - \frac{1}{p^2}\left(\sum_{j=1}^{p}\frac{a_{ja}^2}{h_j^2}\right)^2$$

obteniendo B de manera que sea máxima:

$$V = p^2 S^2 = \sum_{a=1}^{m} p^2 S_a^2(j) = p\sum_{j=1}^{p}\left(\frac{a_{ja}^2}{h_j^2}\right)^2 - \frac{1}{p^2}\left(\sum_{j=1}^{p}\frac{a_{ja}^2}{h_j^2}\right)^2$$

El método Varimax es el más utilizado. El programa SPSS hace por omisión esta

rotación y aplica la normalización de Kaiser.

Otro método de rotación ortogonal es el método Quartimax, pretende a través de simplificar las filas de la matriz de factores, que cada variable tenga una saturación alta con muy pocos factores y próxima a cero con los demás.

El criterio a utilizar es hacer máxima la suma de las cuartas potencias de todas las saturaciones:

$$Q = \sum_{j=1}^{p} \sum_{a=1}^{n} a_{ja}^{4}$$

con la restricción de que la comunalidad de cada variable se ha de mantener constante.

Si **T** es la matriz ortogonal de transformación y **B=AT**, las comunalidades permanecen fijas:

$$\sum_{a=1}^{n} b_{ja}^{2} = \sum_{a=1}^{n} a_{ja}^{2} = h_{j}^{2}$$

Si elevamos al cuadrado esta expresión y sumando las p variables tendremos:

$$\sum_{a=1}^{m} \sum_{j=1}^{p} b_{ja}^{4} + 2 \sum_{i<a}^{m} \sum_{j=1}^{p} b_{ja}^{2} b_{ji}^{2} = Cte.$$

maximizar esta expresión implica minimizar:

$$H = \sum_{i<a}^{n} \sum_{j=1}^{p} b_{ja}^{2} b_{ji}^{2}$$

lo que introduce una estructura más simple de **B**.

Tanto al maximizar **Q** como minimizar **H** obtenemos la misma matriz **T** de transformación.

2.2.11.b.- Rotación Oblicua.

En la rotación oblicua las ponderaciones factoriales no coinciden con las correlaciones entre el factor y la variable puesto que los factores están correlacionados entre si. Pero eso cuando hacemos rotación oblicua la matriz factorial no rotada se convierte en dos matrices diferentes: la matriz de ponderaciones (que es la que se utiliza en la interpretación) y la matriz de correlaciones entre factores y variables.

La perdida de la restauración de ortogonalidad a la matriz de transformación T implica la no incorrelación de los factores. Sin embargo el objetivo será establecer una mejor asociación de cada una de las variables con el factor correspondiente.

2.3. MUESTREO ALEATORIO SIMPLE

2.3.1. INTRODUCCION

Para el presente estudio se desea analizar que factores explican mejor a afectan en mayor grado la compra de los productos en el Canal Tradicional en la ciudad de Guayaquil. Por esta razón se dirigió un cuestionario al detallista (tendero) para que nos comente las tendencias de compra. Dado que se trata de un grupo homogéneo se decidió hacer un muestreo aleatorio simple y es por esta razón que a continuación de detallarán algunas definiciones usadas para el desarrollo de esta parte de la tesis.

Población. Conjunto de individuos o elementos que le podemos observar, medir una característica o atributo.

Ejemplos de población:

- El conjunto formado por todos los estudiantes universitarios en la ESPOL.
- El conjunto de personas fumadoras de una región.

Son características medibles u observables de cada elemento por ejemplo, su estatura, su peso, edad, sexo, etc.

Supongamos que nos interesa conocer el peso promedio de la población formada por los estudiantes de una universidad. Si la universidad tiene 5376 alumnos, bastaría pesar cada estudiante, sumar los 5376 pesajes y dividirlo por 5376. Pero este proceso puede presenta dificultades dentro de las que podemos mencionar:

- localizar y pesar con precisión cada estudiante:
- escribir todos los datos sin equivocaciones en una lista:
- efectuar los cálculos.

Las dificultades son mayores si en número de elementos de la población es infinito, si los elementos se destruyen, si sufren daños al ser medidos o están muy dispersos, si el costo para realizar el trabajo es muy costoso. Una solución a este problema consiste en medir solo una parte de la población que llamaremos muestra y tomar el peso medio en la muestra como una aproximación del verdadero valor del peso medio de la población.

El tamaño de la población es la cantidad de elementos de esta y el tamaño de la muestra es la cantidad de elementos de la muestra. Las poblaciones pueden

ser finitas e infinitas. Los datos obtenidos de una población pueden contener toda la información que se desee de ella. De lo que se trata es de extraerle esa información a la muestra, es decir a los datos muestrales sacarle toda la información de la población.

La muestra debe obtener toda la información deseada para tener la posibilidad de extraerla, esto sólo se puede lograr con una buena selección de la muestra y un trabajo muy cuidadosos y de alta calidad en la recogida de los datos.

Es bueno señalar que en un momento una población puede ser muestra en una investigación y una muestra puede ser población, esto esta dado por el objetivo del investigación, por ejemplo en el caso de determinar la estatura media de los estudiantes universitarios en Cuba una muestra podía ser escoger algunas universidades del país y realizar el trabajo, si por el contrario se quiere saber la estatura promedio de los estudiantes de una universidad en especifico en Cuba, entonces el conjunto formado por todos los estudiantes de esta universidad sería la población y la muestra estaría dada por los grupos, carreras o años seleccionado para realzar el experimento.

Parámetro: Son las medidas o datos que se obtienen sobre la distribución de probabilidades de la población, tales como la media, la varianza, la proporción, etc.

Estadístico: Los datos o medidas que se obtienen sobre una muestra y por lo tanto una estimación de los parámetros.

Error Muestral, de estimación o estándar: Es la diferencia entre un estadístico y su parámetro correspondiente. Es una medida de la variabilidad de las estimaciones de muestras repetidas en torno al valor de la población, nos da una noción clara de hasta dónde y con qué probabilidad una estimación basada en una muestra se aleja del valor que se hubiera obtenido por medio de un censo completo. Siempre se comete un error, pero la naturaleza de la investigación nos indicará hasta qué medida podemos cometerlo (los resultados se someten a error muestral e intervalos de confianza que varían muestra a muestra). Varía según se calcule al principio o al final. Un estadístico será más preciso en cuanto y tanto su error es más pequeño. Podríamos decir que es la desviación de la distribución muestral de un estadístico y su fiabilidad.

Nivel de Confianza. Probabilidad de que la estimación efectuada se ajuste a la realidad. Cualquier información que queremos recoger está distribuida según una ley de probabilidad (Gauss o Student), así llamamos nivel de confianza a la probabilidad de que el intervalo construido en torno a un estadístico capte el verdadero valor del parámetro.

Cuando una población es más homogénea la varianza es menor y el número de entrevistas necesarias para construir un modelo reducido del universo, o de la población, será más pequeño. Generalmente es un valor desconocido y hay que estimarlo a partir de datos de estudios previos.

Para que los resultados obtenidos de los datos muestrales se puedan extender a la población, la muestra debe ser representativa de la población en lo que se refiere a la característica en estudio, o sea, la distribución de la característica en la muestra debe ser aproximadamente igual a la distribución de la característica en la población.

La representatividad en estadística se logra con el tipo de muestreo adecuado que siempre incluye la aleatoriedad en la selección de los elementos de la población que formaran la muestra. No obstante, tales métodos solo nos garantizan una representatividad muy probable pero no completamente segura.

Después de estos preliminares imprescindibles es posible pasa a tratar algunas de las formas que desde el punto de vista científico se puede extraer una muestra.

Al realizar un muestreo en una población podemos hablar de muestreos probabilísticos y no probabilísticos, en nuestro caso nos referiremos a los

muestreos probabilísticos y dentro del mismo estudiaremos el muestreo aleatorio simple (MAS), como método básico en la estadística, el muestreo estratificado y el muestreo por racimos.

Muestreo aleatorio simple: Es aquel en que cada elemento de la población tiene la misma probabilidad de ser seleccionado para integrar la muestra.

Una muestra simple aleatoria es aquella en que sus elementos son seleccionados mediante el muestreo aleatorio simple.

En la práctica no nos interesa el individuo o elemento de la población seleccionado en general, sino solo una característica que mediremos u observaremos en él y cuyo valor será el valor de una variable aleatoria que en cada individuo o elemento de la población puede tomar un valor que será un elemento de cierto conjunto de valores. De modo que una muestra simple aleatoria $x_1, x_2, \ldots, x_n$ se puede interpretar como un conjunto de valores de n variables aleatorias $X_1, X_2, \ldots, X_n$ independientes, cada una de las cuales tiene la misma distribución que es llamada distribución poblacional.

Existen dos formas de extraer una muestra de una población: con reposición y sin reposición.

Muestreo con reemplazo: Es aquel en que un elemento puede ser seleccionado más de una vez en la muestra para ello se extrae un elemento de la población se observa y se devuelve a la población, por lo que de esta forma se pueden hacer infinitas extracciones de la población aun siendo esta finita.

Muestreo sin reemplazo: No se devuelve los elementos extraídos a la población hasta que no se hallan extraídos todos los elementos de la población que conforman la muestra.

Cuando se hace una muestra probabilística debemos tener en cuenta principalmente dos aspectos:

- El método de selección.
- El tamaño de la muestra

2.3.2.- MÉTODO DE SELECCION

Un procedimiento de extraer una muestra aleatoria de una población finita es el de enumerar todos los elementos que conforman la población, escribir esos números en bolas o papelitos echarlos en un bombo o bolsa mezclarlos bien removiéndolos y sacar uno a uno tantos como lo indique el tamaño de la muestra. En este caso los elementos de la muestra lo constituirán los elementos de la población cuyos número coincidan con los extraídos de la bolsa o bombo.

Otro procedimiento para obtener una muestra de una población ya sea el muestreo con replazo o sin reemplazo es mediante la utilización de la tabla de números aleatorios pero solamente para poblaciones finitas, la utilización de estas tablas puede realizarse de diferentes modos pero en el presente trabajo solo expondremos el que consideramos mas eficiente ya que no se necesita de la búsqueda de una gran cantidad innecesaria de números aleatorios en la tabla, el cual será ejemplificado.

Existen diferentes tablas de números aleatorios nosotros en nuestro trabajo utilizaremos como referencia la tabla de M. G. Kendall y B. Babington Smith que se encuentra en el texto de tablas estadísticas, la misma está constituida por 4 bloques de 1000 números aleatorios dispuestos en 25 filas y 40 columnas.

Veamos como se procede para la utilización de la tabla. Consideremos que se desea extraer de una población de tamaño N una muestra de tamaño n se selecciona el bloque, la fila y la columna de la tabla que se va a comenzar, a partir de esta selección (que la hace el muestrista) se toman tantas columnas como dígitos tiene N. Comenzando por el primer número de las columnas seleccionadas se irán incluyendo en la muestra aquellos individuos que en la lista de la población (ya sea de forma horizontal o vertical) ocupa la posición de los n números de las columnas seleccionadas que resultan menores que N, en

los caso que al seleccionar un número en la tabla de números aleatorios sea mayor que N se divide este por N y el resto de la división que será un número entre 0 y N-1 será la posición del individuo a seleccionar tomando el convenio de que el resto 0 corresponde a la posición N. Para la aplicación de este procedimiento requiere que se fije previamente el mayor múltiplo de N que se considerará, para así garantizar que todos los restos desde 0 a N-1 tengan la misma probabilidad de ser seleccionados, por ejemplo si N = 150 y tomando 3 columnas se consideraran sólo aquellos números menores o iguales que 900, los números mayores que 900 no serán analizados en la selección de la muestra.

2.3.3.- EL TAMAÑO DE LA MUESTRA:

Al realizar un muestreo probabilística nos debemos preguntar ¿Cuál es el número mínimo de unidades de análisis (personas, organizaciones, capitulo de telenovelas, etc), que se necesitan para conformar una muestra (n) que me asegure un error estándar menor que 0.01 (fijado por el muestrista o investigador), dado que la población N es aproximadamente de tantos elementos.

En el tamaño de una muestra de una población tenemos que tener presente además si es conocida o no la varianza poblacional.

Para determinar el tamaño de muestra necesario para estimar μ con un error máximo permisible d prefijado y conocida la varianza poblacional (σ^2) podemos utilizar la formula:

$$n = \left(\frac{\sigma Z_{1-\frac{\alpha}{2}}}{d} \right)^2 \quad (1)$$

que se obtiene de reconocer que d es el error estándar o error máximo prefijado y está dado por la expresión $d = \frac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}}$ para el nivel de confianza 1-α y constituye una medida de la precisión de la estimación, por lo que podemos inferir además que $P\left\{\left|\bar{x} - \mu\right| < d\right\} = 1 - \alpha$.

Si la varianza de la población es desconocida, que es lo que mas frecuente se ve en la práctica el tratamiento será diferente, no es posible encontrar una fórmula cuando la varianza poblacional es desconocida por lo que para ello aconsejamos utilizar el siguiente procedimiento:

Primeramente, se toma una pequeña muestra, que se le llama muestra piloto, con ella se estima la varianza poblacional (σ^2) y con este valor se evalúa en la formula (1), sustituyendo (σ^2) por su estimación (S^2). El valor de n obtenido será aproximadamente el valor necesario, nuevamente con ese valor de n se extrae

una muestra de este tamaño de la población se le determina la varianza a esa muestra, como una segunda estimación de (σ^2) y se aplica de nuevo la formula (1), tomando la muestra con el n obtenido como muestra piloto para la siguiente iteración, se llegará a cumplir con las restricciones prefijadas. Se puede plantear esta afirmación ya que la S^2 de σ^2 tiende a estabilizarse a medida que aumenta n alrededor de la σ^2 por lo que llegará el momento en que se encuentre el tamaño de muestra conveniente, sin embargo, en la práctica es mucho más sencillo pues, a lo sumo con tres iteraciones se obtiene el tamaño de muestra deseado, este procedimiento para obtener el tamaño de muestra deseado se puede realizar utilizando en Microsoft Excel en la opción análisis de datos las opciones estadística descriptiva para ir hallando la varianza de cada una de las muestras y la opción muestra para ir determinado las muestras pilotos. Para obtener el tamaño de la muestra utilizando este método recomendamos la utilización de un paquete de computo como por ejemplo el Microsoft Excel, aplicando las opciones muestra y estadística descriptiva.

Para determinar el tamaño de la muestra cuando los datos son cualitativos es decir para el análisis de fenómenos sociales o cuando se utilizan escalas nominales para verificar la ausencia o presencia del fenómeno a estudiar, se recomienda la utilización de la siguiente formula:

$$n = \frac{n'}{1 + \frac{n'}{N}} \qquad (2)$$

siendo $n' = \frac{S^2}{\sigma^2}$ sabiendo que:

σ^2 es la varianza de la población respecto a determinadas variables.

S^2 es la varianza de la muestra, la cual podrá determinarse en términos de probabilidad como $S^2 = p(1-p)$

se es error estándar, que está dado por la diferencia entre $\left(\mu - \bar{x}\right)$ la media poblacional y la media muestral.

$(se)^2$ es el error estándar al cuadrado, que nos servirá para determinar σ^2, por lo que $\sigma^2 = (se)^2$ es la varianza poblacional.

CAPÍTULO III

3. ANALISIS ESTADISTICO

Como ya se mencionó en el CAPITULO I en la metodología, se desarrollarán dos estudios importantes en el presente estudio; el Análisis de series de Tiempo y el de Componentes Principales. Antes de entrar en estos análisis se vio necesario un análisis descriptivo univariante para entender mejor el mercado analizado en el presente estudio.

3.1 ANÁLISIS DESCRIPTIVO UNIVARIANTE

3.1.1 X_1: CERDO

Tabla 3.1.1.a.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Resultados Analisis Descriptivo - Cerdo

N	151
Media	$ 100.656,19
Mediana	$ 98.851,75
Varianza	$ 1.094.328.300,63
CV	32,86%
Q1	$ 79.551,40
Q3	$ 116.093,37

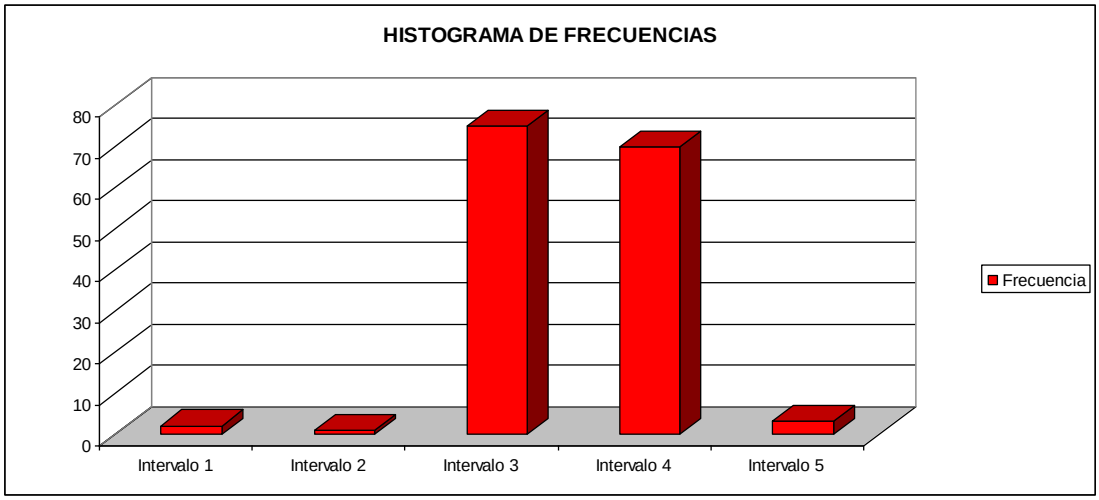
Máximo	$ 225.090,75
Mínimo	$ 19.605,10
Intervalo sugerido basado en número de observaciones	$ 24.941,93

	Valor inferior	Valor Superirior	Frecuencia
Intervalo 1	0	$ 24.941,93	2
Intervalo 2	$ 24.941,93	$ 49.883,87	1
Intervalo 3	$ 49.883,87	$ 99.767,73	75
Intervalo 4	$ 99.767,73	$ 199.535,47	70
Intervalo 5	$ 199.535,47	$ 399.070,93	3

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Tabla 3.1.1.a.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"



6.1

Ver Apéndice A

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Según la investigación de mercado en la pregunta: De cada 10 consumidores ¿Cuántos son mujeres y cuantos son varones?, se obtuvieron los siguientes resultados cualitativos:

Tabla 3.1.1.b.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Analisis de Paretto - Cerdo

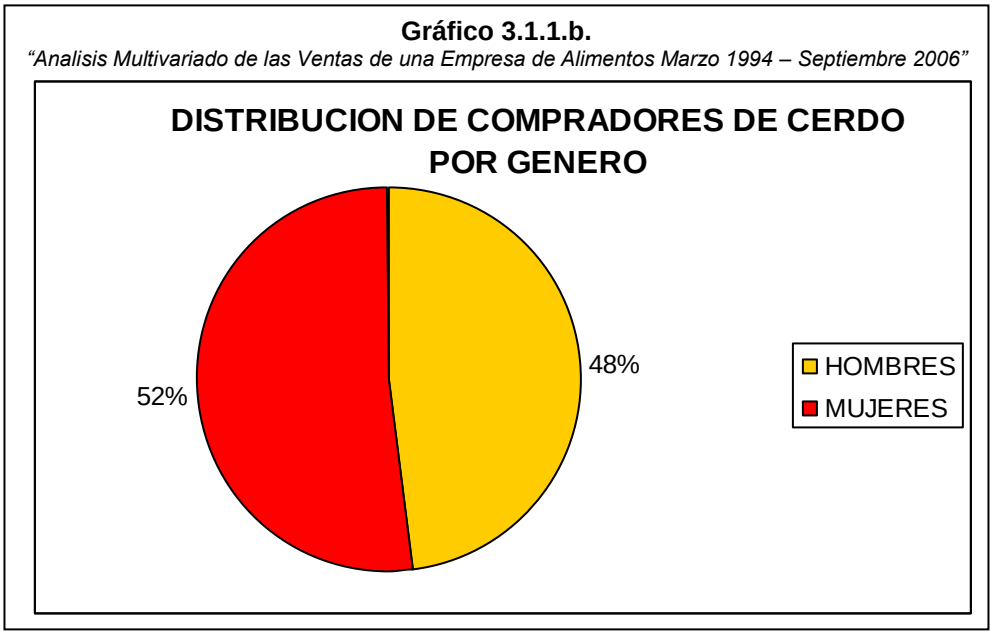
HOMBRES	MUJERES	FRECUENCIA	DIST%
8	2	52	26,3%
0	10	41	20,7%
5	5	31	15,7%
4	6	26	13,1%
7	3	18	9,1%
6	4	11	5,6%
3	7	6	3,0%
2	8	5	2,5%
9	1	3	1,5%
1	9	3	1,5%
10	0	2	1,0%

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Tomando el 80% de los compradores de cerdo podemos que se distribuyen según su género de la siguiente forma:

Gráfico 3.1.1.b.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"



Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Es decir que aunque el género femenino es el mayoritario la diferencia no es representativa. Esto le ayudaría a la empresa en campañas de mercadeo en que usualmente son dirigidas a las amas de casa, y que ahora pueden también comenzar a atacar al género masculino.

De acuerdo a los rangos de edad podemos afirmar lo siguiente:

Tabla 3.1.1.c.
"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Distribución por Rango de Edad

RANGO DE EDAD	FRECUENCIA	FREC.REL.
NO SABE	40	20,20%
MENOR A 15 AÑOS	4	2,02%
ENTRE 15 Y 30 AÑOS	92	46,46%
MAYOR 30 ANOS	62	31,31%

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

La mayoría de los compradores de cerdo en el Canal Tradicional tienen entre quince y treinta años.

3.1.2 X_2: EMBUTIDOS

Tabla 3.1.2.a.
"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Resultados Analisis Descriptivo - Embutidos

N	151
Media	$ 22.434,05
Mediana	$ 21.651,69
Varianza	$ 42.581.107,97
CV	29,09%
Q1	$ 19.728,16
Q3	$ 24.606,14

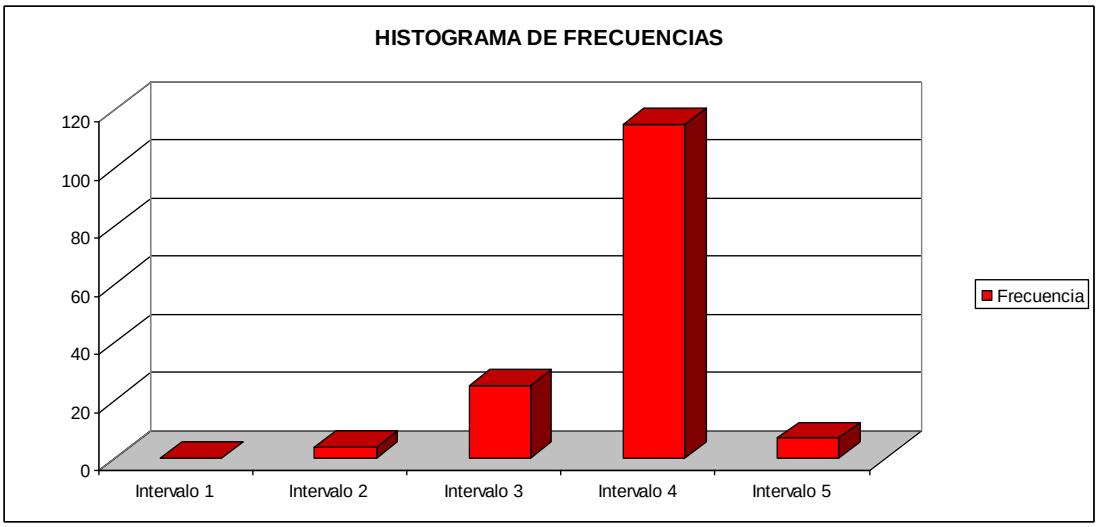
Máximo	$ 45.052,82
Mínimo	$ 6.566,45
Intervalo sugerido basado en número de observaciones	$ 4.671,49

	Valor inferior	Valor Superiri	Frecuencia
Intervalo 1	0	$ 4.671,49	0
Intervalo 2	$ 4.671,49	$ 9.342,98	4
Intervalo 3	$ 9.342,98	$ 18.685,96	25
Intervalo 4	$ 18.685,96	$ 37.371,93	115
Intervalo 5	$ 37.371,93	$ 74.743,86	7

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Tabla 3.1.2.a.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"



Ver Apéndice A

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

En el análisis cualitativo en la línea de embutidos, en la misma pregunta de distribución por género, se presenta descriptivamente de la siguiente manera:

Tabla 3.1.2.b.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Analisis de Paretto - Embutidos

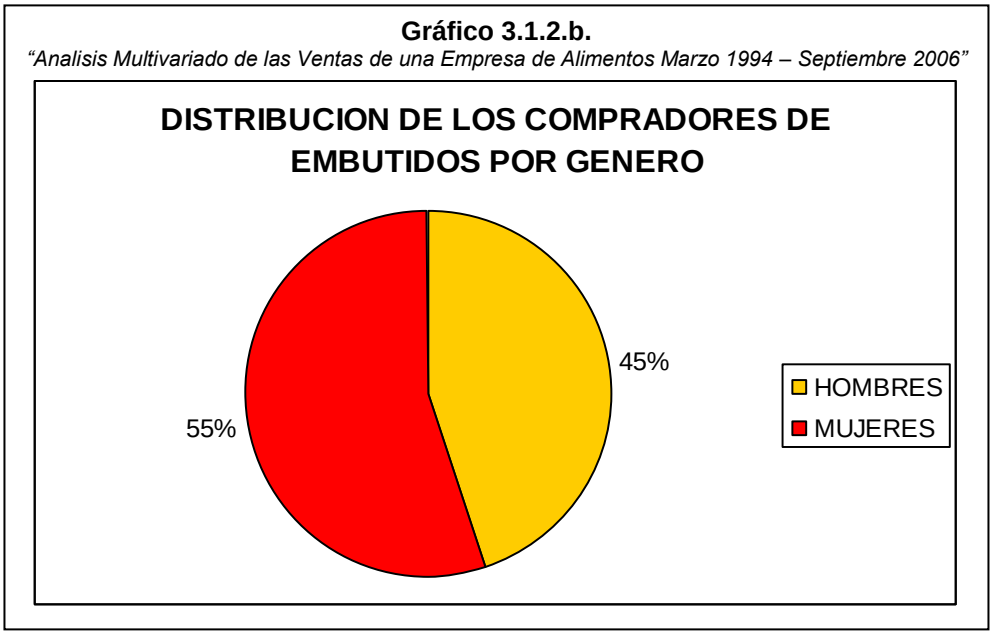
HOMBRES	MUJERES	FRECUENCIA	DIST%
8	2	63	31,82%
5	5	34	17,17%
0	10	25	12,63%
7	3	19	9,60%
4	6	15	7,58%
3	7	13	6,57%
6	4	13	6,57%
2	8	5	2,53%
10	0	5	2,53%
1	9	3	1,52%
9	1	3	1,52%

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Según la encuesta del 80% de los compradores de embutidos en el Canal Tradicional tenemos la siguiente distribución según el género:

Gráfico 3.1.2.b.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"



Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

En cuanto al rango de edad al que pertenecen los compradores de embutidos según los minoristas encuestados tiene el siguiente patrón:

Tabla 3.1.2.c.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Distribución por Rango de Edad

RANGO DE EDAD	FRECUENCIA	FREC.REL.
NO SABE	36	18,18%
MENOR A 15 AÑOS	5	2,53%
ENTRE 15 Y 30 AÑOS	92	46,46%
MAYOR 30 ANOS	65	32,83%

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

La mayoría de los compradores, según la encuesta realizada se encuentran entre quince y treinta años de edad.

3.1.3 X_3: CONGELADOS

Tabla 3.1.3.a.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Resultados Analisis Descriptivo - Embutidos

N	151
Media	$ 2.315,09
Mediana	$ 2.232,95
Varianza	$ 713.644,48
CV	36,49%
Q1	$ 1.796,81
Q3	$ 2.640,84

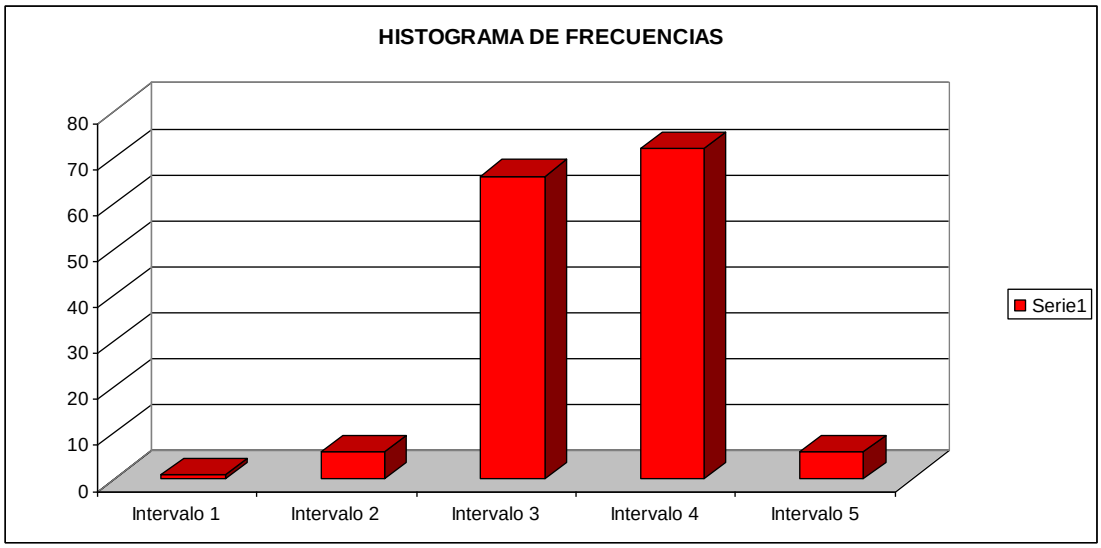
Máximo	$ 4.950,66
Mínimo	$ 470,60
Intervalo sugerido basado en número de observaciones	$ 543,79

	Valor inferior	Valor Superiri	Frecuencia
Intervalo 1	0	$ 543,79	1
Intervalo 2	$ 543,79	$ 1.087,58	6
Intervalo 3	$ 1.087,58	$ 2.175,17	66
Intervalo 4	$ 2.175,17	$ 4.350,33	72
Intervalo 5	$ 4.350,33	$ 8.700,66	6

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Tabla 3.1.3.a.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"



Ver Apéndice A

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

En los productos congelados, según la encuesta de distribución de género, presentan la siguiente información descriptiva:

Tabla 3.1.3.b.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Analisis de Paretto - Congelados

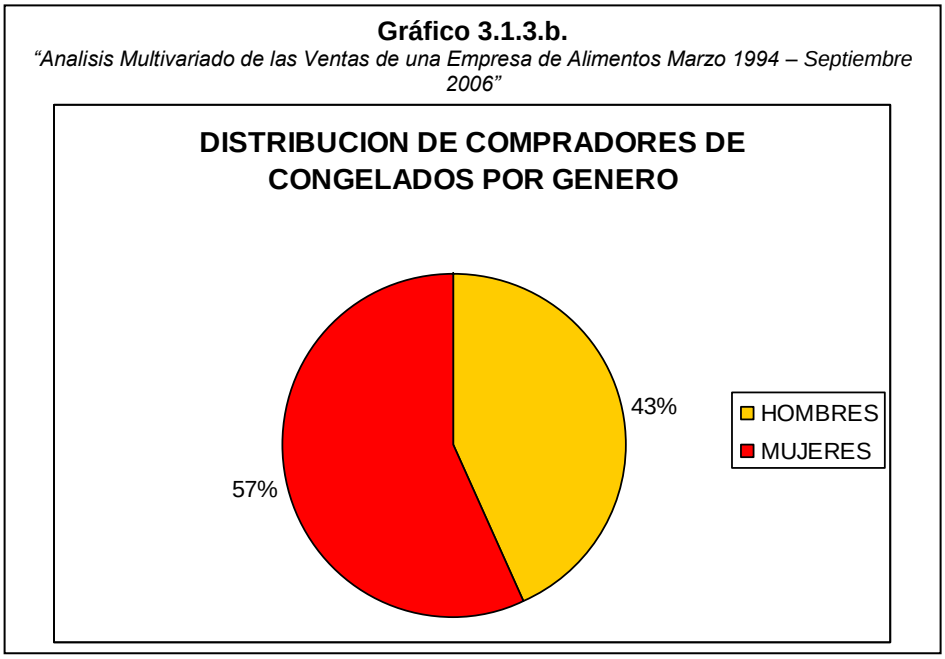
HOMBRES	MUJERES	FRECUENCIA	DIST%
0	10	94	47,47%
8	2	48	24,24%
5	5	31	15,66%
3	7	6	3,03%
4	6	6	3,03%
2	8	5	2,53%
1	9	3	1,52%
7	3	3	1,52%
9	1	1	0,51%
10	0	1	0,51%

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

El 80% de los compradores se distribuyen según el género de la siguiente manera:

Gráfico 3.1.3.b.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"



Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

El rango de edad en que se encuentran los compradores de productos congelados, según la investigación de mercado sigue la siguiente distribución:

Tabla 3.1.2.c.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Distribución por Rango de Edad

RANGO DE EDAD	FRECUENCIA	FREC.REL.
NO SABE	88	44,44%
MENOR A 15 AÑOS	2	1,01%
ENTRE 15 Y 30 AÑOS	49	24,75%
MAYOR 30 ANOS	59	29,80%

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

La mayoría de los compradores productos congelados se encuentran en el rango de mayores a treinta años.

3.1.4 X_4: PRODUCTOS DEL MAR

Tabla 3.1.4.a.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Resultados Analisis Descriptivo – Productos del Mar

N	151
Media	$ 4.854,61
Mediana	$ 4.690,21
Varianza	$ 6.233.727,54
CV	51,43%
Q1	$ 3.371,39
Q3	$ 6.216,12

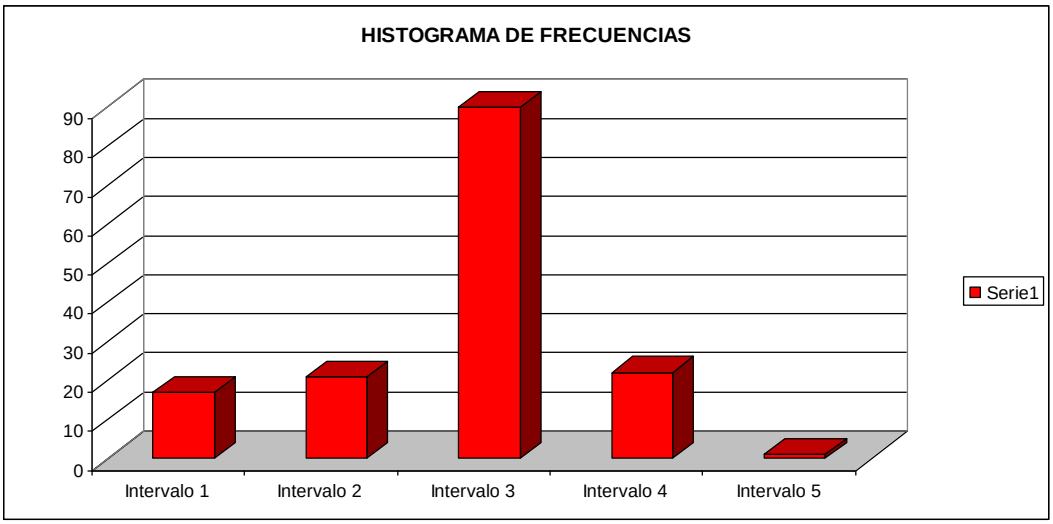
Máximo	$ 14.054,80
Mínimo	$ -
Intervalo sugerido basado en número de observaciones	$ 1.705,98

	Valor inferior	Valor Superiric	Frecuencia
Intervalo 1	0	$ 1.705,98	17
Intervalo 2	$ 1.705,98	$ 3.411,95	21
Intervalo 3	$ 3.411,95	$ 6.823,91	90
Intervalo 4	$ 6.823,91	$ 13.647,82	22
Intervalo 5	$ 13.647,82	$ 27.295,64	1

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Tabla 3.1.4.a.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"



Ver Apéndice A

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

El comportamiento de la línea de Productos del Mar se comporta muy parecido a la línea de Congelados siguiendo las siguientes características en cuanto a la distribución de género:

Tabla 3.1.4.b.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Analisis de Paretto – Productos del Mar

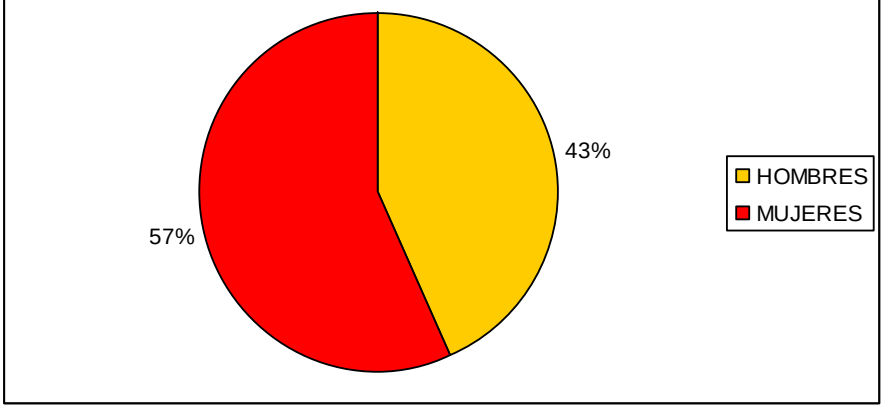
HOMBRES	MUJERES	FRECUENCIA	DIST%
0	10	81	40,91%
8	2	49	24,75%
5	5	32	16,16%
3	7	10	5,05%
6	4	7	3,54%
7	3	5	2,53%
1	9	4	2,02%
4	6	4	2,02%
2	8	2	1,01%
9	1	2	1,01%
10	0	2	1,01%

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Con una distribución por género:

Gráfico 3.1.4.b.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"



Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Y con la mayoría de sus compradores en el rango de mayores a 30 años.

Tabla 3.1.4.c.
"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Distribución por Rango de Edad

RANGO DE EDAD	FRECUENCIA	FREC.REL.
NO SABE	89	44,95%
MENOR A 15 AÑOS	2	1,01%
ENTRE 15 Y 30 AÑOS	49	24,75%
MAYOR 30 ANOS	58	29,29%

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

3.1.5 X_5: CONSERVAS

Tabla 3.1.5.a.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Resultados Analisis Descriptivo – Conservas

N	151
Media	$ 101.881,91
Mediana	$ 96.007,80
Varianza	$ 1.918.028.155,52
CV	42,99%
Q1	$ 72.565,36
Q3	$ 124.588,53

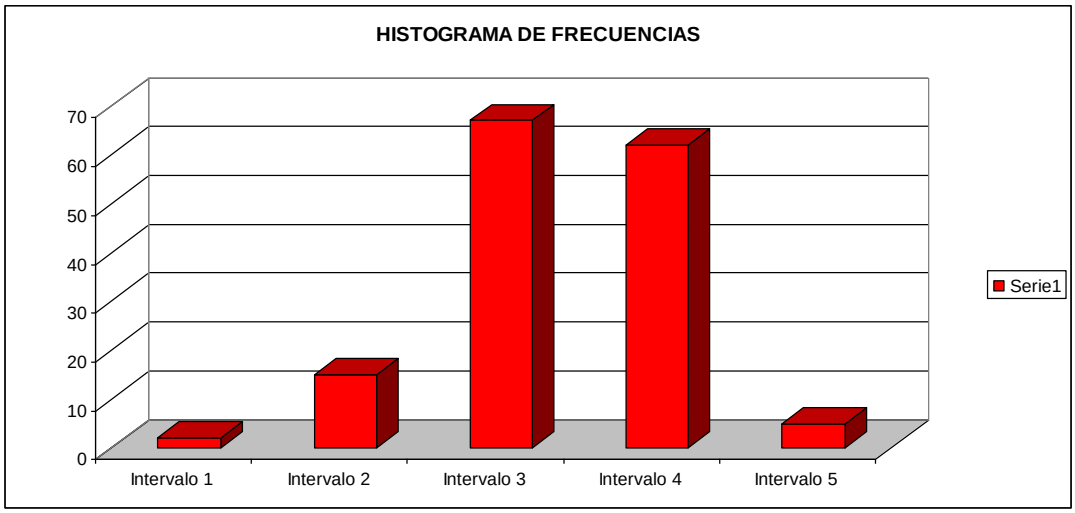
Máximo	$ 230.606,95
Mínimo	$ 19.400,40
Intervalo sugerido basado en número de observaciones	$ 25.636,34

	Valor inferior	Valor Superirior	Frecuencia
Intervalo 1	0	$ 25.636,34	2
Intervalo 2	$ 25.636,34	$ 51.272,68	15
Intervalo 3	$ 51.272,68	$ 102.545,35	67
Intervalo 4	$ 102.545,35	$ 205.090,70	62
Intervalo 5	$ 205.090,70	$ 410.181,41	5

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Tabla 3.1.5.a.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"



Ver Apéndice A

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

La línea de conservas cualitativamente hablando se presenta descriptivamente de la siguiente manera:

Del 80% de sus compradores se presenta el género femenino como mayoría. La diferencia es un poco más pronunciada.

Tabla 3.1.5.b.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

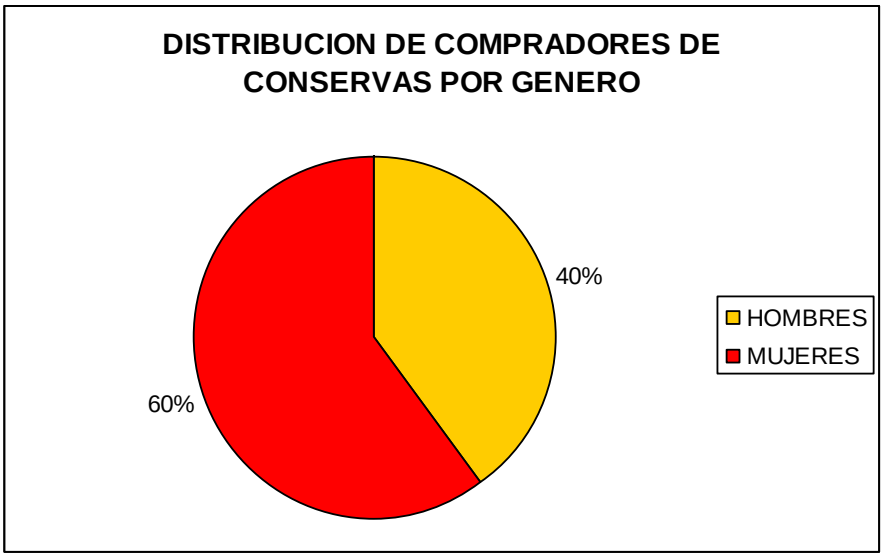
Analisis de Paretto – Conservas

HOMBRES	MUJERES	FRECUENCIA	DIST%
0	10	44	22,22%
8	2	51	25,76%
5	5	37	18,69%
3	7	22	11,11%
4	6	16	8,08%
2	8	9	4,55%
7	3	8	4,04%
6	4	5	2,53%
1	9	3	1,52%
9	1	3	1,52%

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Gráfico 3.1.5.b.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"



Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

El rango de edad que contiene a la mayoría de compradores entre quince y treinta años:

Tabla 3.1.5.c.
"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Distribución por Rango de Edad

RANGO DE EDAD	FRECUENCIA	FREC.REL.
NO SABE	52	26,26%
MENOR A 15 AÑOS	8	4,04%
ENTRE 15 Y 30 AÑOS	74	37,37%
MAYOR 30 ANOS	64	32,32%

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

3.1.6 X_6: ARROZ

Tabla 3.1.6.a.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Resultados Analisis Descriptivo – Arroz

N	151
Media	$ 28.850,04
Mediana	$ 26.324,07
Varianza	$ 234.417.145,17
CV	53,07%
Q1	$ 17.069,73
Q3	$ 36.613,33

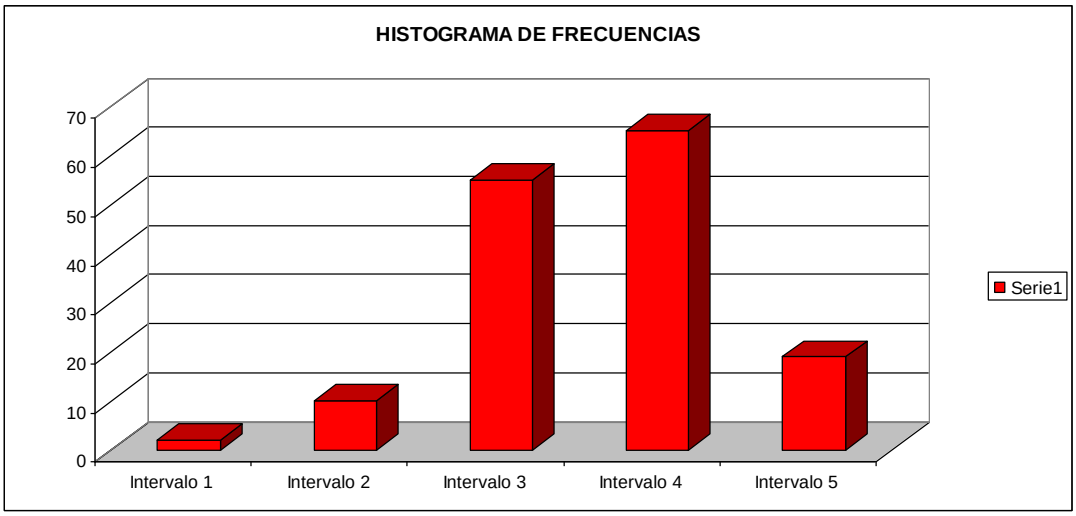
Máximo	$ 83.755,71
Mínimo	$ 34.196,81
Intervalo sugerido basado en número de observaciones	$ 6.015,48

	Valor inferior	Valor Superirior	Frecuencia
Intervalo 1	0	$ 6.015,48	2
Intervalo 2	$ 6.015,48	$ 12.030,96	10
Intervalo 3	$ 12.030,96	$ 24.061,92	55
Intervalo 4	$ 24.061,92	$ 48.123,84	65
Intervalo 5	$ 48.123,84	$ 96.247,68	19

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Tabla 3.1.6.a.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"



Ver Apéndice A

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

La línea de arroz presenta la siguiente información descriptiva en la investigación cualitativa:

Tabla 3.1.6.b.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Analisis de Paretto – Arroz

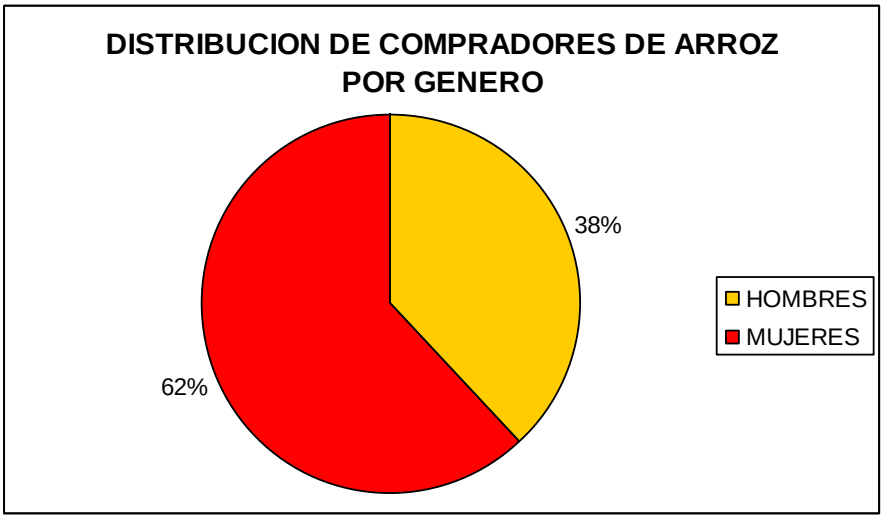
HOMBRES	MUJERES	FRECUENCIA	DIST%
0	10	56	28,43%
8	2	53	26,90%
5	5	35	17,77%
4	6	13	6,60%
2	8	10	5,08%
6	4	7	3,55%
3	7	6	3,05%
7	3	6	3,05%
10	0	5	2,54%
1	9	3	1,52%
9	1	3	1,52%

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Con el 80% de los compradores de arroz se obtuvo la siguiente distribución según el género:

Gráfico 3.1.6.b.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"



Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

El rango de edad con el mayor porcentaje de los compradores es el de mayores a treinta años:

Tabla 3.1.6.c.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Distribución por Rango de Edad

RANGO DE EDAD	FRECUENCIA	FREC.REL.
NO SABE	59	29,80%
MENOR A 15 AÑOS	2	1,01%
ENTRE 15 Y 30 AÑOS	65	32,83%
MAYOR 30 ANOS	72	36,36%

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

3.2. ANÁLISIS DE SERIES DE TIEMPO

Se obtuvo la información de ventas por línea desde marzo de 1994 hasta septiembre del año 2006. Los datos obtenidos se muestran en el Apéndice A. Para el análisis de series de tiempo se utilizara el programa Demetra, el cual estima modelos en base a métodos de modelación ARIMA. Los modelos a continuación pasaron todas las pruebas de evaluación de modelos según Demetra.

3.2.1. CERDO

La línea de cerdo presento la siguiente estimación del modelo de series de tiempo:

Tabla 3.2.1.a.
"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Modelo de Ventas Mensuales - Cerdo

Information on Models	Model 1 (Tramo-Seats)
Specif. of the ARIMA model	(0 1 1)(0 1 1) (fixed)
Non-seas. MA (lag 1) value	-0.6866
Non-seas. MA (lag 1) t-value	-11.10 [-1.972, 1.972] 5%
Seasonal MA (lag 12) value	-0.9874
Seasonal MA (lag 12) t-value	-73.19 [-1.972, 1.972] 5%

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Como se puede observar los parámetros estimados del modelo: $\Delta\Delta_{12}y_t = (1+0{,}9874L^{12})(1+0{,}6866L)\xi_t$ son significativos.

De acuerdo a este modelo se obtuvo la siguiente proyección:

Tabla 3.2.1.b.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Proyección Estimada Vs. Venta Real

Mes	Proyeccion	Venta Real
Oct-06	$ 122.720,92	$ 122.719,98
Nov-06	$ 147.866,91	$ 147.867,00
Dic-06	$ 182.340,53	$ 182.370,74
Ene-07	$ 109.674,44	
Feb-07	$ 116.836,21	
Mar-07	$ 116.003,75	
Abr-07	$ 125.487,58	
May-07	$ 144.703,69	
Jun-07	$ 126.368,81	
Jul-07	$ 119.237,62	
Ago-07	$ 129.897,80	
Sep-07	$ 119.197,11	
Oct-07	$ 126.754,49	
Nov-07	$ 154.716,83	
Dic-07	$ 174.611,56	
Ene-08	$ 108.900,11	
Feb-08	$ 133.562,65	
Mar-08	$ 117.111,85	
Abr-08	$ 128.047,41	
May-08	$ 148.587,86	
Jun-08	$ 128.067,33	
Jul-08	$ 119.864,39	
Ago-08	$ 130.397,38	
Sep-08	$ 120.766,98	

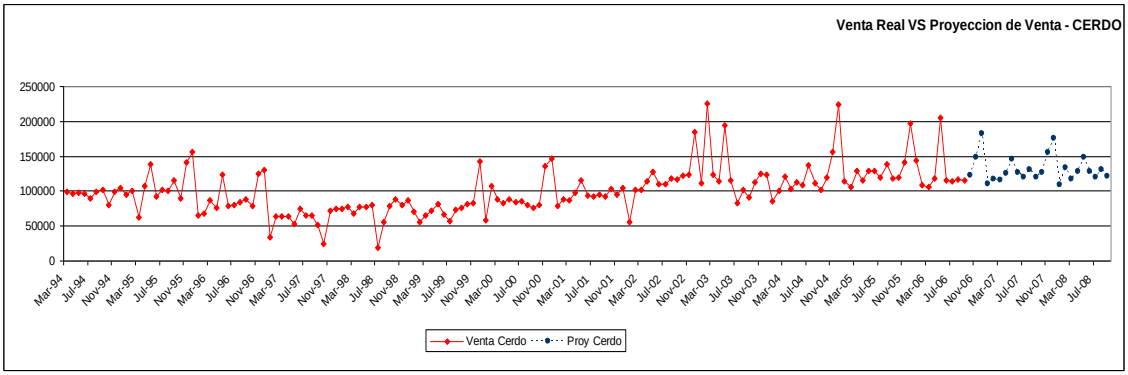
Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

El gráfico a continuación muestra la proyección a partir de la venta real.

Gráfico 3.2.1

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Pronóstico de la Venta



Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

3.2.2 EMBUTIDOS

La línea de embutidos presento el siguiente modelo con coeficientes significativos:

Tabla 3.2.2.a.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Modelo de Ventas Mensuales – Embutidos

Information on Models	Model 1 (Tramo-Seats)
Specif. of the ARIMA model	(0 1 1)(0 1 1) (fixed)
Non-seas. MA (lag 1) value	-0.7492
Non-seas. MA (lag 1) t-value	-12.18 [-1.972, 1.972] 5%
Seasonal MA (lag 12) value	-0.8436
Seasonal MA (lag 12) t-value	-12.00 [-1.972, 1.972] 5%

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Con este modelo: $\Delta\Delta_{12}y_t = (1+0{,}8436L^{12})(1+0{,}7492L)\xi_t$ se obtuvo la siguiente proyección:

Tabla 3.2.2.b.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Proyección Estimada Vs. Venta Real

Mes	Proyeccion	Venta Real
Oct-06	$ 22.538,28	$ 22.538,70
Nov-06	$ 26.622,11	$ 26.621,97
Dic-06	$ 33.680,94	$ 33.681,27
Ene-07	$ 20.766,50	
Feb-07	$ 23.028,84	
Mar-07	$ 22.748,66	
Abr-07	$ 21.686,49	
May-07	$ 27.017,19	
Jun-07	$ 23.679,49	
Jul-07	$ 22.435,48	
Ago-07	$ 23.043,53	
Sep-07	$ 21.983,17	
Oct-07	$ 21.315,62	
Nov-07	$ 28.580,62	
Dic-07	$ 32.225,85	
Ene-08	$ 22.032,20	
Feb-08	$ 27.215,04	
Mar-08	$ 21.210,12	
Abr-08	$ 22.896,75	
May-08	$ 27.417,12	
Jun-08	$ 22.416,56	
Jul-08	$ 22.362,77	
Ago-08	$ 24.222,09	
Sep-08	$ 22.591,01	

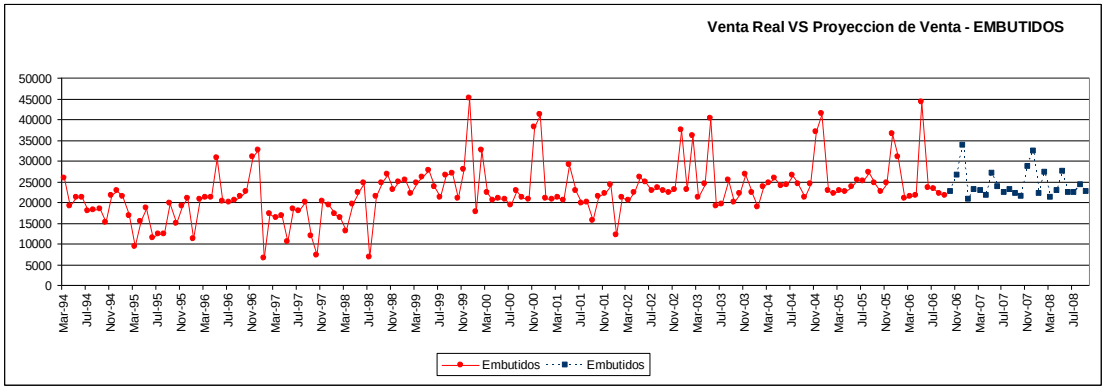
Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

El siguiente grafico muestra la proyección en base de la venta real:

Gráfico 3.2.2

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Pronóstico de la Venta



Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

3.2.3. CONGELADOS

Los productos congelados presentaron el siguiente modelo: $\Delta\Delta_{12}y_t = (1+0{,}9874L^{12})(1+0{,}6866L)\xi_t$ con coeficientes significativos según Demetra:

Tabla 3.2.3.a.
"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Modelo de Ventas Mensuales - Congelados

Information on Models	Model 1 (Tramo-Seats)
Specif. of the ARIMA model	(0 1 1)(0 1 1) (fixed)
Non-seas. MA (lag 1) value	-0.7313
Non-seas. MA (lag 1) t-value	-12.46 [-1.972, 1.972] 5%
Seasonal MA (lag 12) value	-0.9842
Seasonal MA (lag 12) t-value	-64.64 [-1.972, 1.972] 5%

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Este modelo nos ayuda a estimar los siguientes pronósticos:

Tabla 3.2.3.b.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Proyección Estimada Vs. Venta Real

Mes	Proyeccion	Venta Real
Oct-06	$ 2.387,58	$ 2.387,52
Nov-06	$ 2.990,93	$ 2.991,07
Dic-06	$ 3.570,00	$ 3.569,83
Ene-07	$ 2.435,38	
Feb-07	$ 2.671,95	
Mar-07	$ 2.496,68	
Abr-07	$ 2.730,64	
May-07	$ 3.189,40	
Jun-07	$ 2.588,75	
Jul-07	$ 2.539,93	
Ago-07	$ 2.667,23	
Sep-07	$ 2.537,29	
Oct-07	$ 2.495,34	
Nov-07	$ 3.089,35	
Dic-07	$ 3.668,43	
Ene-08	$ 2.533,81	
Feb-08	$ 2.773,05	
Mar-08	$ 2.585,77	
Abr-08	$ 2.838,40	
May-08	$ 3.278,49	
Jun-08	$ 2.687,17	
Jul-08	$ 2.647,68	
Ago-08	$ 2.746,99	
Sep-08	$ 2.654,38	

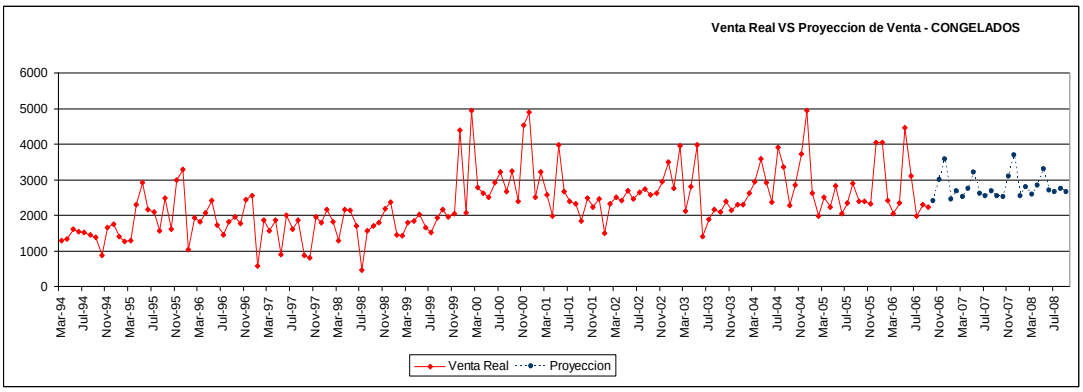
Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Esta proyección se ilustra a partir de la venta real en el siguiente gráfico:

Gráfico 3.2.3

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Pronóstico de la Venta



Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

3.2.4. PRODUCTOS DEL MAR

En la línea de productos del mar se obtiene el siguiente modelo con coeficientes significativos: $\Delta\Delta_{12}y_t = (1+0{,}9874L^{12})(1+0{,}6866L)\xi_t$

Tabla 3.2.4.a.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Modelo de Ventas Mensuales - Congelados

Information on Models	Model 1 (Tramo-Seats)
Specif. of the ARIMA model	(0 1 1)(0 1 1) (fixed)
Non-seas. MA (lag 1) value	-0.7057
Non-seas. MA (lag 1) t-value	-11.40 [-1.972, 1.972] 5%
Seasonal MA (lag 12) value	-0.9494
Seasonal MA (lag 12) t-value	-34.61 [-1.972, 1.972] 5%

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Del cual se obtiene los siguientes pronósticos:

Tabla 3.2.4.b.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

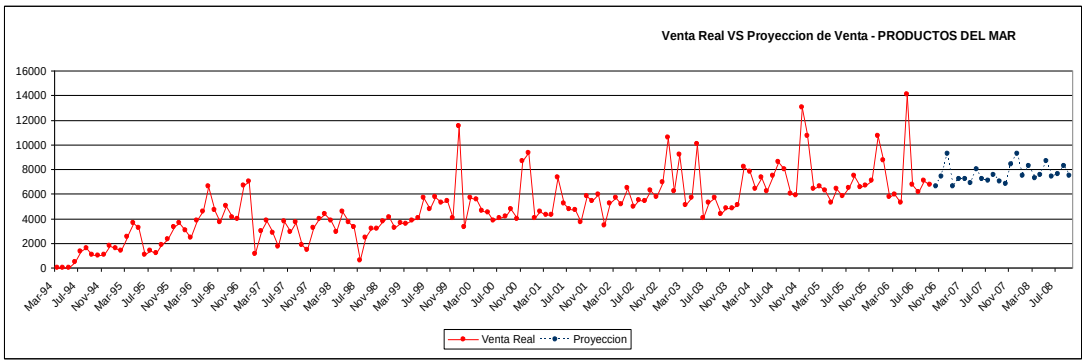
Proyección Estimada Vs. Venta Real

Mes	Proyeccion	Venta Real
Oct-06	$ 6.599,94	$ 6.601,00
Nov-06	$ 7.394,93	$ 7.395,01
Dic-06	$ 9.279,67	$ 9.281,23
Ene-07	$ 6.608,56	
Feb-07	$ 7.212,57	
Mar-07	$ 7.196,80	
Abr-07	$ 6.858,74	
May-07	$ 7.988,19	
Jun-07	$ 7.223,99	
Jul-07	$ 7.069,69	
Ago-07	$ 7.558,19	
Sep-07	$ 6.979,28	
Oct-07	$ 6.790,81	
Nov-07	$ 8.389,64	
Dic-07	$ 9.256,07	
Ene-08	$ 7.449,20	
Feb-08	$ 8.247,61	
Mar-08	$ 7.248,71	
Abr-08	$ 7.565,91	
May-08	$ 8.646,69	
Jun-08	$ 7.416,93	
Jul-08	$ 7.581,67	
Ago-08	$ 8.265,36	
Sep-08	$ 7.438,41	

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Con el siguiente gráfico podemos observar esta proyección a partir de la venta real:

Gráfico 3.2.3
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Pronóstico de la Venta



Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

3.2.5. CONSERVAS

La línea de conservas presenta un modelo más diferente al de los productos carnicol:

Tabla 3.2.5.a.
"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Modelo de Ventas Mensuales - Congelados

Information on Models	Model 1 (Tramo-Seats)
Specif. of the ARIMA model	(0 1 2)(0 0 0) (fixed)
Non-seas. MA (lag 1) value	-0.7457
Non-seas. MA (lag 1) t-value	-8.91 [-1.972, 1.972] 5%
Non-seas. MA (lag 2) value	-0.2236
Non-seas. MA (lag 2) t-value	-2.65 [-1.972, 1.972] 5%

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Este MA(2) se lo expresaría de la siguiente forma: $\Delta y_t = (1 + 0{,}2236L + 0{,}2236L^2)\xi_t$. Los pronósticos a continuación se obtuvieron del modelo presentado:

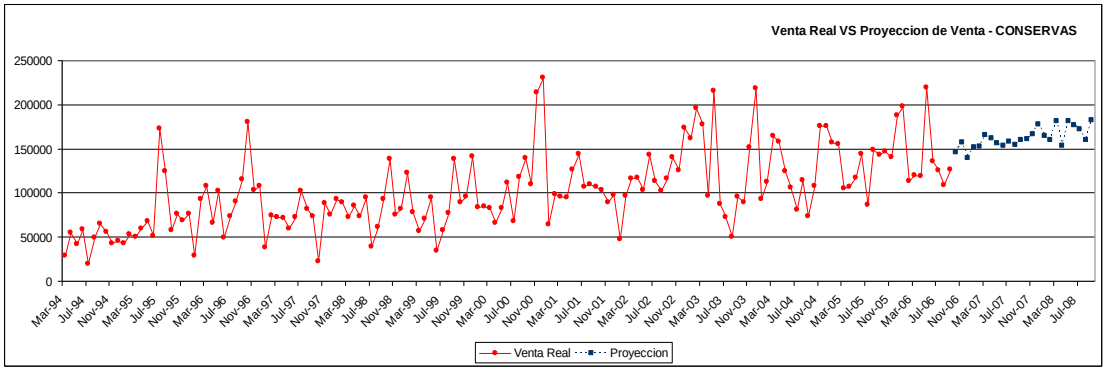
Tabla 3.2.5.b.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Proyección Estimada Vs. Venta Real

Mes	Proyeccion	Venta Real
Oct-06	$ 146.028,28	$ 146.030,32
Nov-06	$ 157.038,93	$ 156.998,15
Dic-06	$ 139.771,08	$ 140.005,50
Ene-07	$ 151.353,52	
Feb-07	$ 152.353,46	
Mar-07	$ 165.177,67	
Abr-07	$ 161.331,85	
May-07	$ 156.422,86	
Jun-07	$ 153.405,87	
Jul-07	$ 157.549,04	
Ago-07	$ 154.652,55	
Sep-07	$ 159.827,33	
Oct-07	$ 160.614,33	
Nov-07	$ 166.499,62	
Dic-07	$ 177.873,08	
Ene-08	$ 164.902,12	
Feb-08	$ 159.520,86	
Mar-08	$ 181.429,32	
Abr-08	$ 153.149,96	
May-08	$ 181.163,98	
Jun-08	$ 176.945,94	
Jul-08	$ 171.561,86	
Ago-08	$ 159.489,58	
Sep-08	$ 182.291,51	

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

En el gráfico a continuación se puede observar este pronóstico a partir de las ventas reales:

Gráfico 3.2.5
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Pronóstico de la Venta



Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

3.2.6. ARROZ

Por ultimo el modelo para la línea de arroz es el siguiente:

$$\Delta\Delta_{12} y_t = (1 + 0{,}9623L^{12})(1 + 0{,}9464L)\xi_t$$

Tabla 3.2.6.a.

"Analisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Modelo de Ventas Mensuales - Congelados

Information on Models	Model 1 (Tramo-Seats)
Specif. of the ARIMA model	(0 1 1)(0 1 1) (fixed)
Non-seas. MA (lag 1) value	-0.9464
Non-seas. MA (lag 1) t-value	-34.00 [-1.972, 1.972] 5%
Seasonal MA (lag 12) value	-0.9623
Seasonal MA (lag 12) t-value	-41.07 [-1.972, 1.972] 5%

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

El presente modelo arrojó los siguientes pronósticos:

Tabla 3.2.6.b.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Proyección Estimada Vs. Venta Real

Mes	Proyeccion	Venta Real
Oct-06	$ 24.572,64	$ 24.602,73
Nov-06	$ 26.556,46	$ 26.560,00
Dic-06	$ 26.851,78	$ 27.200,41
Ene-07	$ 20.734,61	
Feb-07	$ 22.002,34	
Mar-07	$ 20.053,32	
Abr-07	$ 20.318,29	
May-07	$ 22.957,03	
Jun-07	$ 27.539,96	
Jul-07	$ 21.597,48	
Ago-07	$ 23.715,47	
Sep-07	$ 19.039,24	
Oct-07	$ 19.623,07	
Nov-07	$ 21.155,84	
Dic-07	$ 21.339,19	
Ene-08	$ 16.437,87	
Feb-08	$ 17.400,57	
Mar-08	$ 15.820,71	
Abr-08	$ 15.990,85	
May-08	$ 18.023,75	
Jun-08	$ 21.569,37	
Jul-08	$ 16.874,16	
Ago-08	$ 18.483,99	
Sep-08	$ 14.803,30	

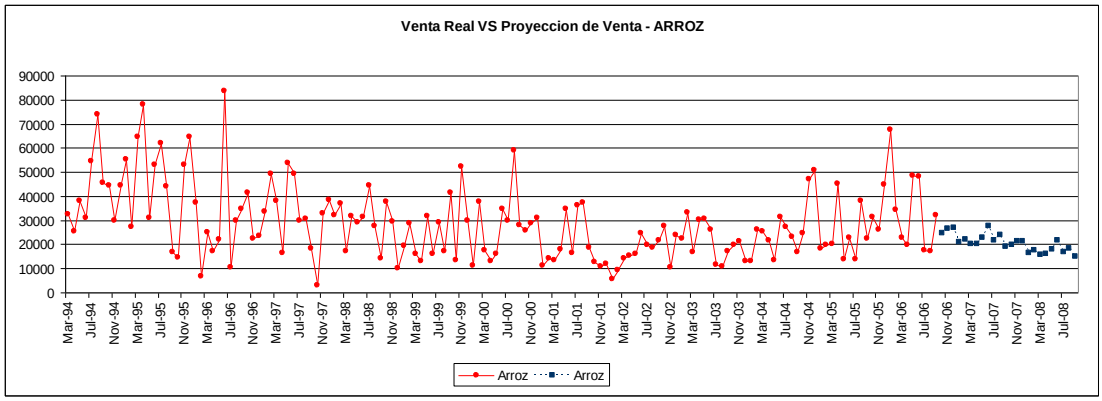
Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

A partir de las ventas reales tenemos el pronóstico en el siguiente gráfico:

Gráfico 3.2.6

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Pronóstico de la Venta



Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

3.3. ANALISIS DE COMPONENTES PRINCIPALES

Antes de entrar en el análisis de componentes principales es necesario indicar que la encuesta realizada tanto para los resultados cualitativos descriptivos como para el presente análisis fue en base un muestreo aleatorio simple utilizando un error del 0.15 a un 95% de nivel de confianza.

Partiendo de esta base se obtuvo un n_0=170 con un tamaño de población N=1514 se obtiene un tamaño de muestra n=153. Se tomaron 198 observaciones de la población.

En el estudio a continuación se desea determinar cuales son los factores de consumo que influyen en la compra con mayor fuerza por lo que es necesario que se desarrolle un Análisis de Componentes Principales.

Se presentaron 10 características por las cuales el consumidor elige un producto, para que sean calificados (0-Menos Importante a 5 Más Importante) en cada línea. Las variables o características que influyen en la compra a analizar son:

Tabla 3.3.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Variables Originales

VARIABLE	NOMBRE
Y1	CALIDAD
Y2	SABOR
Y3	PRECIO
Y4	EMPAQUE
Y5	COSTUMBRE
Y6	FRESCURA
Y7	TEXTURA
Y8	PROMOCIONES
Y9	IMPULSO
Y10	PUBLICIDAD

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Se utilizó MINITAB 13.20 para procesar la información para luego interpretarla.

3.3.1. CERDO

Según el Análisis de Componentes Principales se obtienen los siguientes resultados:

Tabla 3.3.1.a.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Eiegenanalisis de la Matriz de Covarianzas CERDO										
Eigenvalue	44,702	2,649	1,216	0,467	0,351	0,237	0,166	0,115	0,06	0,048
Proportion	0,894	0,053	0,024	0,009	0,007	0,005	0,003	0,002	0,001	0,001
Cumulative	0,894	0,947	0,971	0,98	0,987	0,992	0,996	0,998	0,999	1

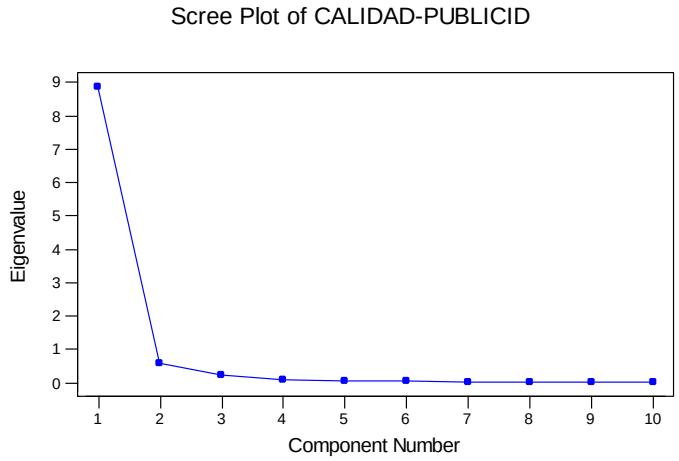
Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
CALIDAD	-0,325	0,007	-0,478	0,312	0,401	0,093	-0,122	-0,066	-0,596	0,156
SABOR	-0,327	0,019	-0,42	0,197	0,263	-0,006	-0,026	-0,012	0,727	-0,281
TEXTURA	-0,33	-0,068	-0,224	-0,092	-0,496	-0,421	0,26	-0,105	-0,263	-0,507
EMPAQUE	-0,331	-0,054	-0,215	-0,164	-0,311	-0,254	0,067	0,175	0,184	0,765
COSTUMBR	-0,327	-0,079	0,029	-0,369	-0,193	0,295	-0,73	0,241	-0,047	-0,178
FRESCURA	-0,332	-0,007	0,048	-0,224	-0,073	0,726	0,542	-0,102	-0,007	0,01
PRECIO	-0,211	0,956	0,174	0,06	-0,05	-0,056	-0,037	0,013	-0,018	0,004
PROMOCIO	-0,323	-0,114	0,371	-0,232	0,515	-0,295	0,234	0,518	-0,076	-0,102
IMPULSO	-0,323	-0,121	0,353	-0,176	0,209	-0,196	-0,161	-0,776	0,056	0,111
PUBLICID	-0,314	-0,211	0,45	0,744	-0,272	0,086	-0,055	0,122	0,031	0,024

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Gráfico 3.3.1.a.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

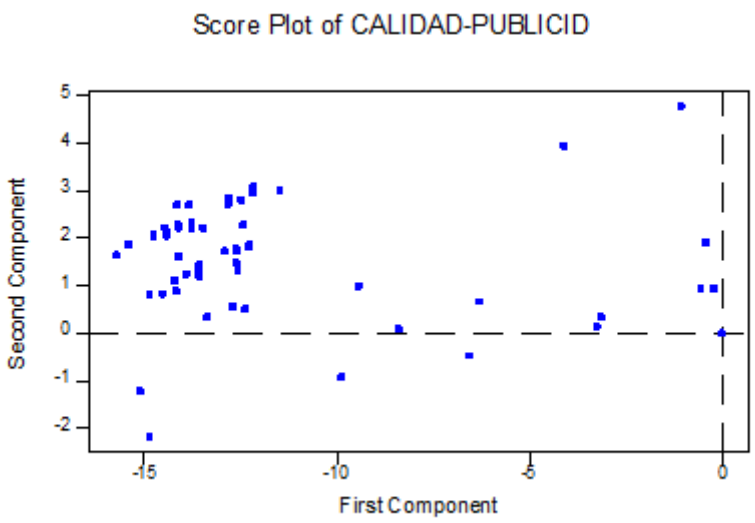
Eigenvalores Vs. Número de Componentes



Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Como se puede observar el 89% de la varianza lo explica lo explica el primer componente.

Gráfico 3.3.1.b.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Nube de Puntos



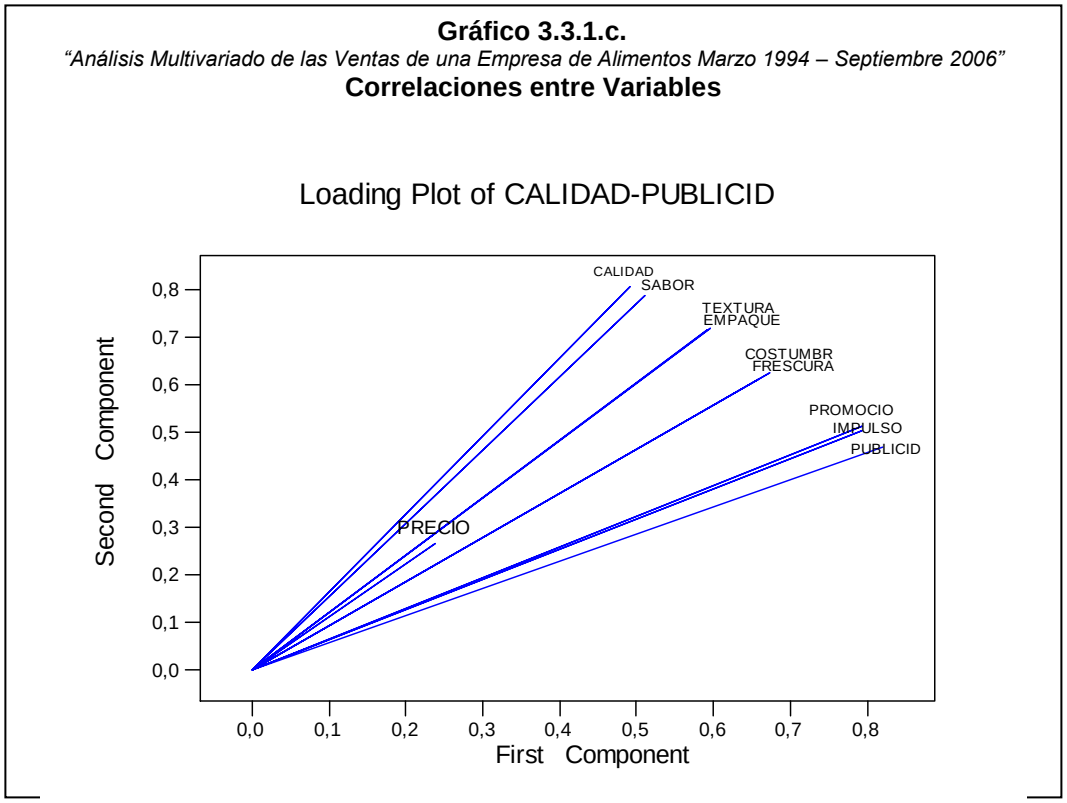
Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

En la primera componente principal (que explica el 89,4% de la varianza) todas las variables originales tienen igual peso, por lo que a este eje lo denominaremos "Imagen de Marca". A la segunda componente principal la denominaremos "Eje del Precio", puesto que la variable PRECIO es la que tiene mayor peso.

Gráfico 3.3.1.c.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Correlaciones entre Variables



Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

El coseno del Angulo entre las variables es el coeficiente de correlación entre ellas. De acuerdo al gráfico 3.3.1.c. se puede decir entre las variables PROMOCION, PUBLICIDAD e IMPULSO, existe una alta correlación. Así mismo, formando un grupo aparte las variables COSTUMBRE y FRESCURA. Igual Sucede con TEXTURA y EMPAQUE.

3.3.2. EMBUTIDOS

En el análisis de componentes principales se obtuvieron los siguientes resultados:

Tabla 3.3.2.a.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Eiegenanalisis de la Matriz de Covarianzas EMBUTIDOS										
Eigenvalue	42,408	2,959	1,133	0,588	0,402	0,261	0,175	0,095	0,062	0,024
Proportion	0,882	0,062	0,024	0,012	0,008	0,005	0,004	0,002	0,001	0,001
Cumulative	0,882	0,943	0,967	0,979	0,987	0,993	0,996	0,998	0,999	1

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
CALIDAD	-0,319	0,048	-0,683	0,467	0,396	0,134	0,124	-0,141	-0,021	0,015
SABOR	-0,338	0,077	-0,239	-0,014	-0,213	-0,259	-0,303	0,781	0,059	-0,068
TEXTURA	-0,337	-0,011	-0,111	-0,192	-0,222	-0,233	-0,325	-0,377	-0,104	0,689
EMPAQUE	-0,336	-0,024	-0,067	-0,178	-0,144	-0,229	-0,261	-0,449	0,053	-0,71
COSTUMBR	-0,327	-0,099	0,07	-0,242	0,09	-0,469	0,755	0,059	-0,134	0,021
FRESCURA	-0,332	0,008	-0,112	-0,243	-0,492	0,701	0,288	0,024	0,026	-0,017
PRECIO	-0,195	0,929	0,283	0,1	0,067	0,01	0,04	-0,048	-0,029	0,006
PROMOCIO	-0,322	-0,139	0,268	-0,104	0,368	0,11	-0,056	0,02	0,794	0,115
IMPULSO	-0,321	-0,168	0,28	-0,22	0,503	0,289	-0,248	0,135	-0,57	-0,045
PUBLICID	-0,309	-0,267	0,459	0,726	-0,292	-0,029	0,024	-0,027	-0,091	0

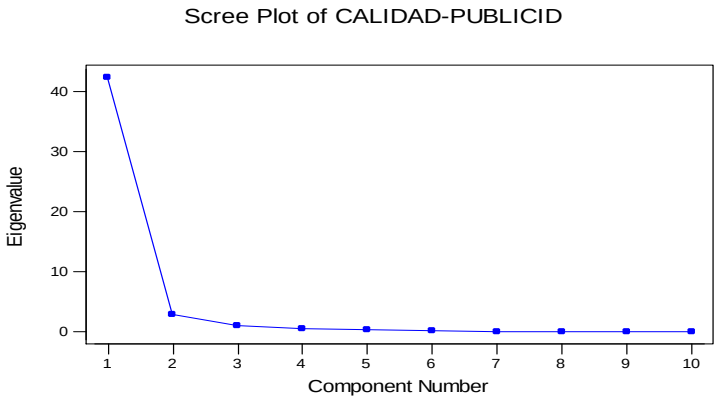
Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

El 88% de la varianza es explicada por el primer componente.

Gráfico 3.3.2.a.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Eigenvalores Vs. Número de Componentes

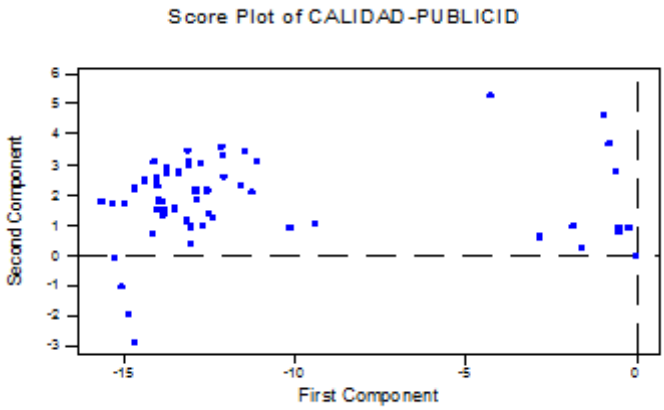


Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Gráfico 3.3.2.b.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Nube de Puntos



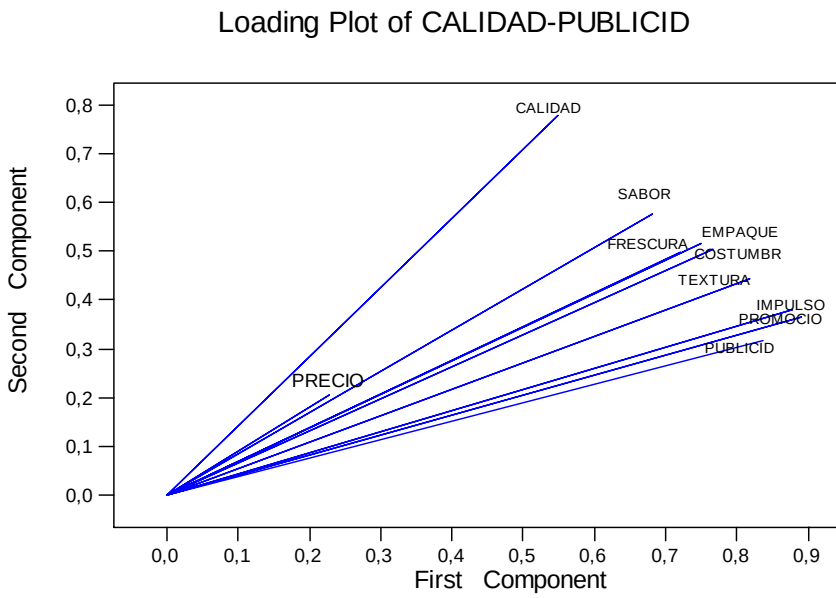
Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

En la primera componente principal, el cual explica 88,2% de la varianza, todas las variables originales poseen pesos similares, por esta razón también denominaremos a este eje, “Imagen de Marca”. A la segunda componente principal, dado que la variable PRECIO vuelve a tener el mayor peso, se lo denominara “Eje del Precio”.

Gráfico 3.3.2.c.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Correlaciones entre Variables



.

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Aunque las correlaciones están un poco menos visibles en el grafico, si se puede DISTINGUIR una correlación mas marcada entre las variables IMPULSO, PROMOCION y PUBLICIDAD y aparte en otro grupo EMPAQUE, COSTUMBRE y FRESCURA. La variable CALIDAD no esta correlacionada con los otras variables en una proporción representativa por lo que se puede concluir que la calidad del producto de embutidos es muy poco importante.

3.3.3. CONGELADOS

El análisis de componentes principales arrojo los siguientes resultados:

Tabla 3.3.3.a.

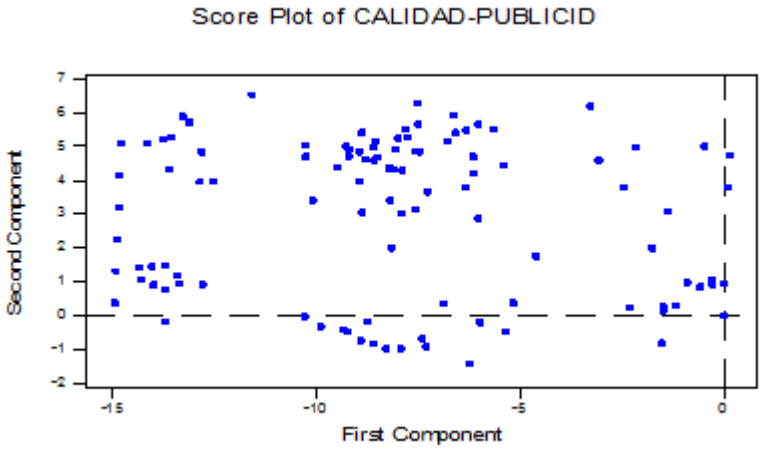
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Eiegenanalisis de la Matriz de Covarianzas CONGELADOS										
Eigenvalue	33,7	4,632	3,734	1,851	1,154	1,117	0,992	0,829	0,683	0,275
Proportion	0,688	0,095	0,076	0,038	0,024	0,023	0,02	0,017	0,014	0,006
Cumulative	0,688	0,783	0,859	0,897	0,92	0,943	0,963	0,98	0,994	1

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
CALIDAD	-0,306	-0,162	0,418	-0,796	-0,132	0,025	0,112	0,175	0,1	0,007
SABOR	-0,34	0,139	0,226	0,096	0,166	-0,565	0,038	-0,578	0,342	-0,072
PRECIO	-0,31	0,135	-0,298	-0,046	-0,436	0,246	0,468	-0,434	-0,268	0,256
EMPAQUE	-0,351	0,161	-0,181	0,09	-0,429	-0,229	-0,52	0,303	0,17	0,427
COSTUMBR	-0,357	-0,023	-0,431	0,028	-0,136	-0,131	0,169	0,293	0,105	-0,724
FRESCURA	-0,303	-0,046	-0,494	-0,255	0,711	0,029	-0,046	0,031	-0,063	0,289
TEXTURA	0,032	0,945	0,124	-0,109	0,141	0,081	0,1	0,193	-0,039	-0,069
PROMOCIO	-0,349	-0,001	0,248	0,071	0,063	0,038	-0,433	-0,135	-0,732	-0,253
IMPULSO	-0,34	-0,005	0,164	0,228	0,094	0,73	-0,189	-0,118	0,453	-0,097
PUBLICID	-0,338	-0,125	0,344	0,46	0,157	-0,094	0,482	0,446	-0,122	0,246

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Gráfico 3.3.3.a.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Eigenvalores Vs. Número de Componentes



Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

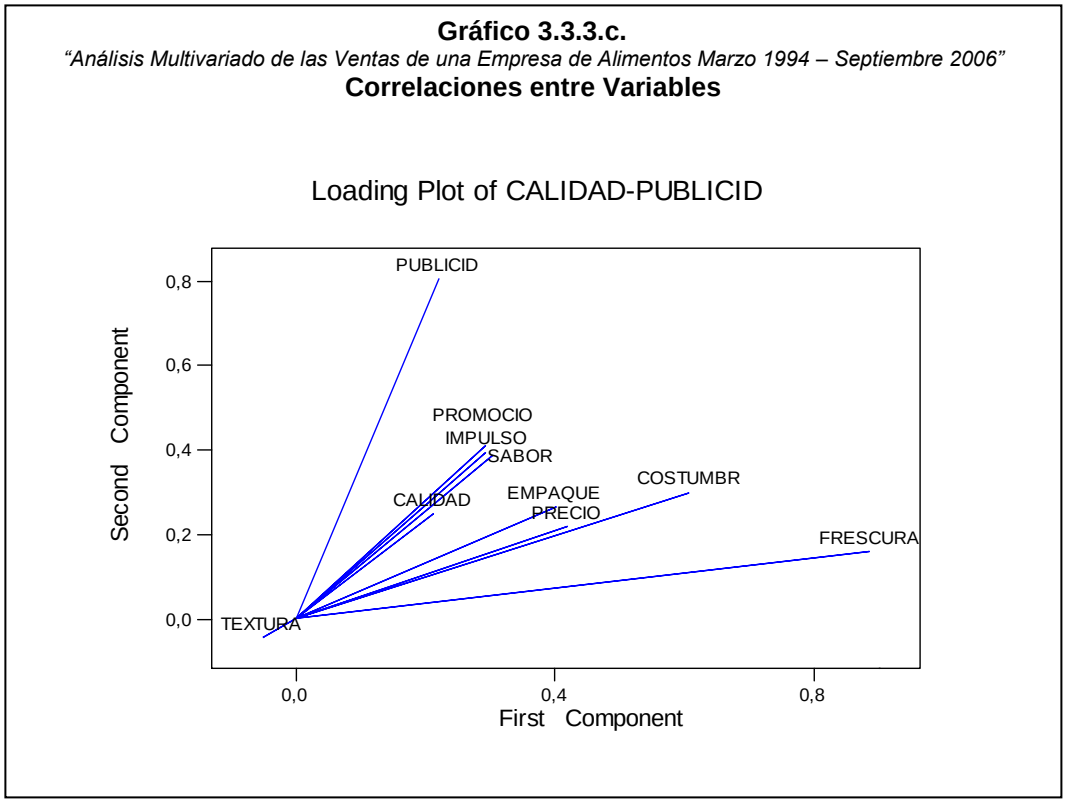
Por la misma razón que en CERDO Y EMBUTIDOS la primera componente(que explica 88,8% de la varianza), también la denominaremos "Imagen de Marca" y a la segunda componente la denominaremos "Eje de la Textura" porque la variable original TEXTURA es la de mayor peso.

La correlación entre las variables se las puede apreciar en el siguiente grafico

Gráfico 3.3.3.c.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Correlaciones entre Variables



Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Se puede observar que las variables SABOR, IMPULSO, PROMOCION y CALIDAD están altamente correlacionadas y puede ser debido a que este producto no posee una venta de volumen en el Canal Tradicional y el detallista calificó a estas características como necesarias junto con PRECIO. También se pueden apreciar que las variables EMPAQUE, PRECIO y COSTUMBRE están altamente correlacionadas.

3.3.4. PRODUCTOS DEL MAR

Para Productos del Mar se obtienen los siguientes resultados en el análisis de componentes principales:

Tabla 3.3.4.a.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

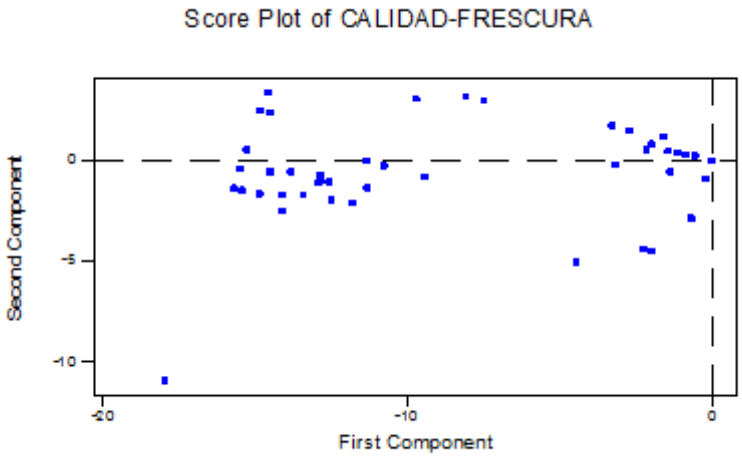
Eiegenanalisis de la Matriz de Covarianzas PRODUCTOS DEL MAR										
Eigenvalue	49,68	2,625	1,508	1,189	0,478	0,356	0,2	0,153	0,079	0,033
Proportion	0,882	0,047	0,027	0,021	0,008	0,006	0,004	0,003	0,001	0,001
Cumulative	0,882	0,929	0,956	0,977	0,985	0,992	0,995	0,998	0,999	1

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
CALIDAD	-0,319	0,222	-0,116	-0,66	-0,608	-0,138	0,059	0,081	0,01	-0,031
SABOR	-0,336	0,111	-0,008	-0,295	0,356	0,303	-0,179	-0,731	0,053	-0,011
PRECIO	-0,287	0,088	-0,864	0,391	-0,076	-0,035	0,005	-0,055	0,015	0,002
EMPAQUE	-0,339	0,091	-0,005	-0,138	0,348	0,174	0,031	0,448	0,006	0,711
COSTUMBR	-0,333	0,105	0,014	-0,089	0,379	0,163	0,158	0,433	0,057	-0,695
TEXTURA	-0,227	-0,951	-0,085	-0,145	-0,071	0,098	0,035	0,016	-0,01	-0,016
PROMOCIO	-0,329	-0,036	0,225	0,19	0,01	-0,485	-0,168	-0,036	0,733	0,024
IMPULSO	-0,323	0,047	0,274	0,293	-0,148	0,02	0,797	-0,234	-0,118	0,08
PUBLICID	-0,318	0,062	0,293	0,386	-0,429	0,508	-0,446	0,102	-0,084	-0,041
FRESCURA	-0,335	-0,018	0,144	0,079	0,146	-0,573	-0,273	-0,028	-0,659	-0,037

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

El 88% de la varianza es explicado por el primer componente principal.

Gráfico 3.3.4.a.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Eigenvalores Vs. Número de Componentes



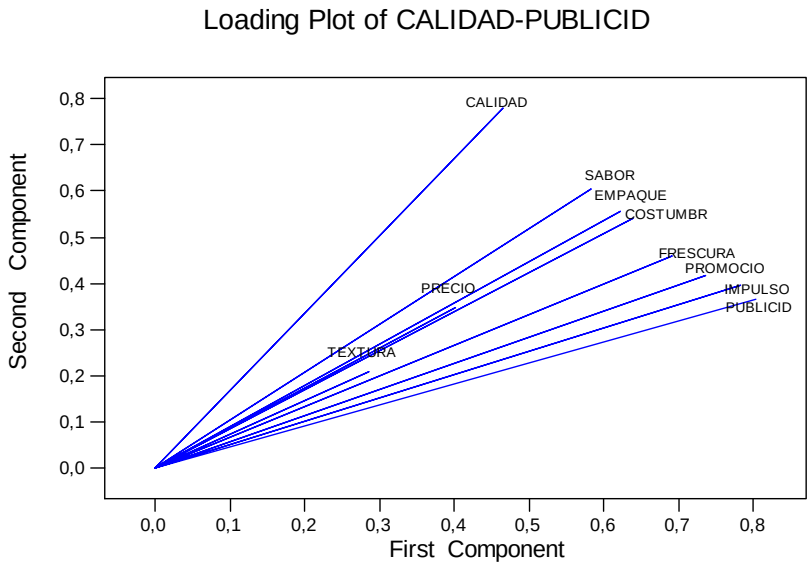
Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

En la primera componente principal, el cual explica 88,2% de la varianza, todas las variables originales poseen pesos similares, por esta razón también denominaremos a este eje, "Imagen de Marca". A la segunda componente principal, dado que la variable CALIDAD tiene el mayor peso, se lo denominara "Eje de la Calidad".

Gráfico 3.3.4.c.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Correlaciones entre Variables



Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Todas las variables están altamente.

3.3.5. CONSERVAS

Según el análisis de componentes principales se obtuvieron los siguientes resultados:

Tabla 3.3.5.a.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Eiegenanalisis de la Matriz de Covarianzas CONSERVAS										
Eigenvalue	48,938	2,912	1,172	0,629	0,366	0,155	0,114	0,048	0,041	0,01
Proportion	0,9	0,054	0,022	0,012	0,007	0,003	0,002	0,001	0,001	0
Cumulative	0,9	0,953	0,975	0,986	0,993	0,996	0,998	0,999	1	1

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
CALIDAD	-0,318	-0,088	-0,716	0,575	0,114	0,114	0,147	0,018	0,02	-0,003
SABOR	-0,336	0,025	-0,18	-0,073	-0,32	-0,429	-0,686	-0,299	-0,004	-0,056
PRECIO	-0,334	0,008	-0,023	-0,18	-0,144	-0,257	0,14	0,509	-0,078	0,695
EMPAQUE	-0,334	0,004	-0,009	-0,198	-0,147	-0,096	0,166	0,478	-0,286	-0,693
COSTUMBR	-0,331	-0,031	0,023	-0,243	-0,062	-0,226	0,503	-0,338	0,631	-0,1
FRESCURA	-0,33	0,022	-0,036	-0,284	-0,383	0,765	-0,013	-0,216	-0,131	0,117
TEXTURA	-0,195	0,923	0,197	0,259	0,047	0,01	0,035	-0,031	-0,001	-0,002
PROMOCIO	-0,327	-0,061	0,109	-0,14	0,604	0,264	-0,424	0,282	0,404	-0,042
IMPULSO	-0,328	-0,105	0,122	-0,13	0,529	-0,145	0,164	-0,425	-0,578	0,085
PUBLICID	-0,304	-0,352	0,622	0,592	-0,213	0,027	-0,009	0,004	0,026	-0,001

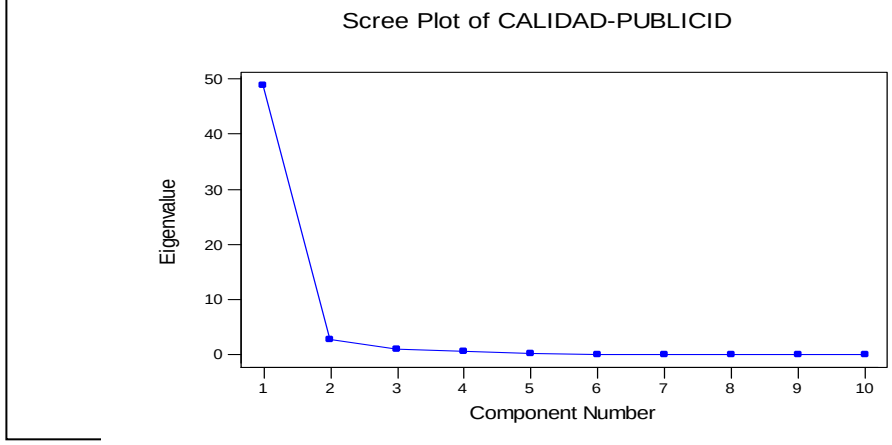
Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

El 90% de la varianza es explicada por el primer componente principal.

Gráfico 3.3.5.a.

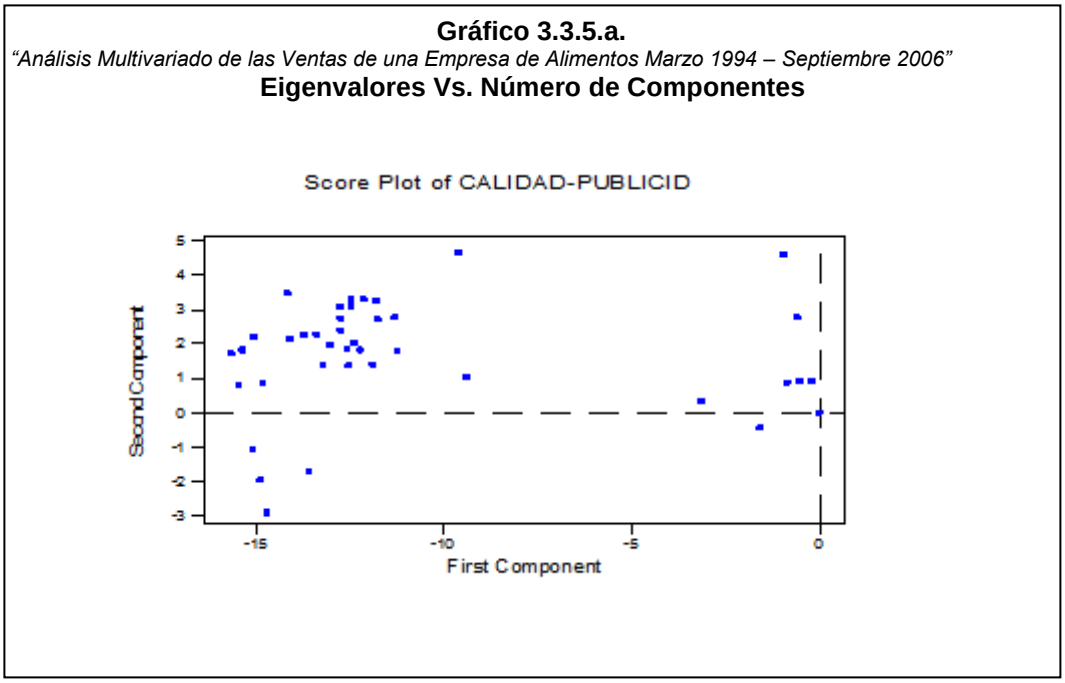
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Eigenvalores Vs. Número de Componentes



Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Gráfico 3.3.5.a.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Eigenvalores Vs. Número de Componentes

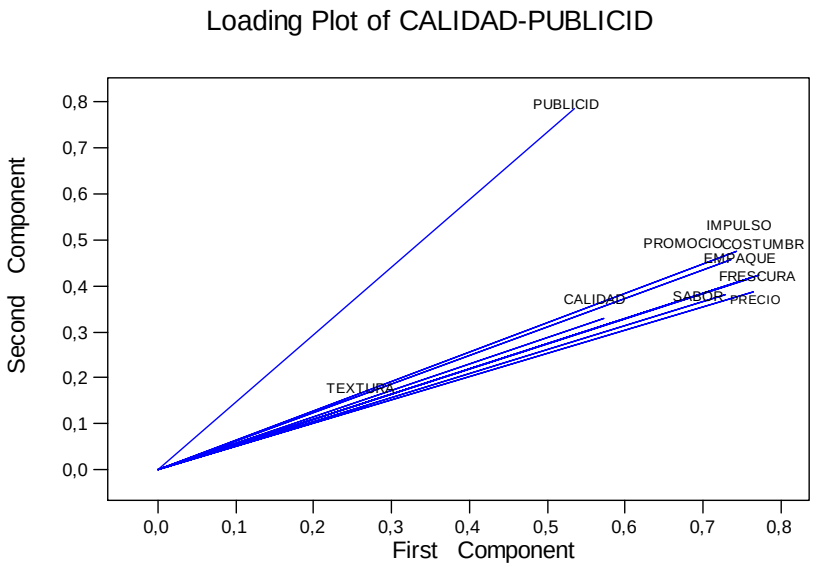


Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

En la primera componente principal, que explica 90% de la varianza, todas las variables originales poseen pesos similares, por esta razón denominaremos a este eje, "Imagen de Marca". A la segunda componente principal, dado que la variable TEXTURA tiene el mayor peso, se lo denominara "Eje de la Textura".

A continuación se puede observar que existe una gran exigencia en cuanto al consumo de las CONSERVAS. Casi todas las variables están fuertemente correlacionadas.

Gráfico 3.3.5.c.
"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"
Correlaciones entre Variables



Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

3.3.6. Arroz

Se obtuvieron los siguientes resultados en el análisis de componentes principales:

Tabla 3.3.6.a.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Eiegenanalisis de la Matriz de Covarianzas ARROZ										
Eigenvalue	51,608	2,51	1,073	0,555	0,277	0,136	0,099	0,093	0,011	0,006
Proportion	0,916	0,045	0,019	0,01	0,005	0,002	0,002	0,002	0	0
Cumulative	0,916	0,96	0,979	0,989	0,994	0,996	0,998	1	1	1

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
CALIDAD	-0,316	-0,156	-0,827	0,363	-0,242	-0,044	-0,007	0,012	-0,006	-0,004
SABOR	-0,335	-0,002	-0,11	-0,267	0,398	-0,014	0,069	0,007	0,113	0,793
PRECIO	-0,334	-0,014	-0,09	-0,27	0,357	-0,074	-0,166	-0,08	0,626	-0,499
EMPAQUE	-0,333	-0,021	-0,063	-0,249	0,34	-0,177	-0,048	-0,053	-0,764	-0,294
COSTUMBR	-0,328	-0,05	0,184	-0,157	-0,435	0,23	-0,728	-0,195	-0,057	0,13
FRESCURA	-0,327	-0,006	0,051	-0,201	-0,165	0,712	0,411	0,356	-0,032	-0,138
TEXTURA	-0,194	0,959	-0,014	0,202	-0,011	-0,013	-0,024	0,001	-0,004	-0,006
PROMOCIO	-0,326	-0,035	0,203	-0,107	-0,395	-0,294	0,515	-0,573	0,061	-0,001
IMPULSO	-0,326	-0,091	0,283	0,091	-0,218	-0,523	-0,015	0,686	0,063	0,018
PUBLICID	-0,316	-0,206	0,367	0,733	0,346	0,199	-0,009	-0,164	-0,005	-0,004

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

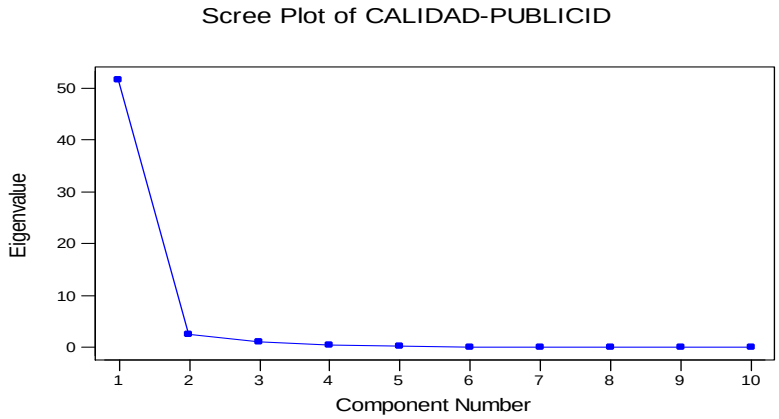
En la primera componente principal, que explica 91,6% de la varianza, todas las variables originales poseen pesos similares, por esta razón denominaremos a este eje, "Imagen de Marca". A la segunda componente principal, dado que la variable TEXTURA tiene el mayor peso, se lo denominara "Eje de la Textura".

La importancia de la textura tiene sentido, dado que el Canal Tradicional está conformado por las tiendas y clientes informales, los consumidores, en cuanto al tema de granos, buscan palpar el producto al granel para saber si esta en buen estado.

Gráfico 3.3.6.a.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Eigenvalores Vs. Número de Componentes

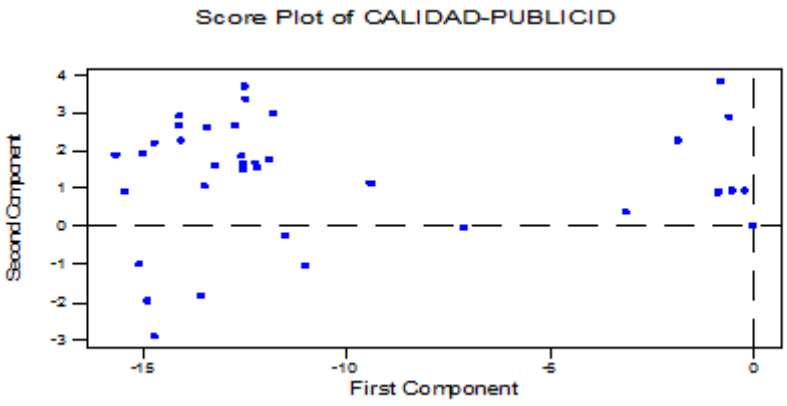


Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Gráfico 3.3.6.a.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Eigenvalores Vs. Número de Componentes



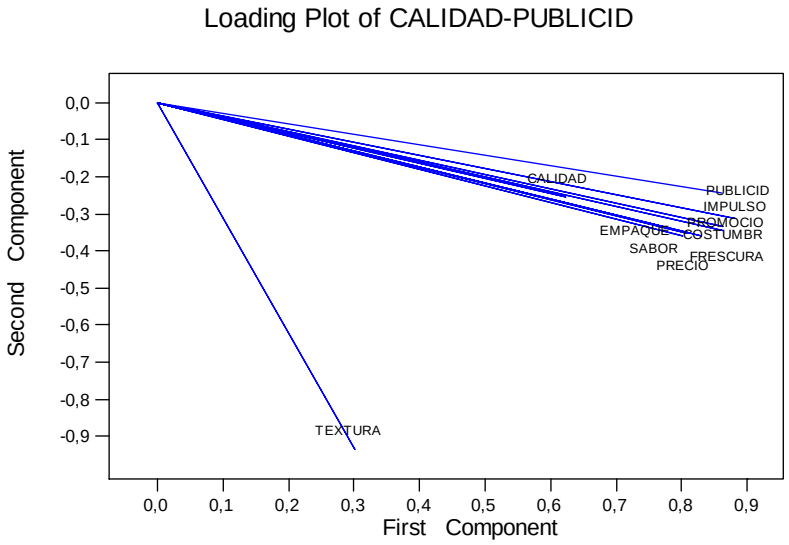
Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Se puede observar como se correlacionan las variables en el siguiente gráfico:

Gráfico 3.3.6.c.

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Correlaciones entre Variables



Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Al igual que las conservas se muestra que las variables se encuentran muy correlacionadas entre si.

CAPITULO IV

4. CONCLUSIONES Y RECOMENDACIONES

4.1 Conclusiones

1. En la distribución por género las variables Cerdo, Embutidos y Congela no existe una diferencia significativa. Por el contrario la tendencia de compra de los Productos del Mar, Conservas, y Arroz si se distinguen mayoritariamente por el género femenino.

2. En cuanto al rango de edad de los consumidores, el estudio nos indica que, en todas las variables, la mayoría pertenece al rango de entre 15 y 30 años de edad.

3. Se obtuvo un modelo de series de tiempo que pasa todas las pruebas para la variable CERDO:

$$\Delta\Delta_{12} y_t = (1 + 0{,}9874L^{12})(1 + 0{,}6866L)\xi_t$$

El pronósticos comparados con la venta real del los meses de octubre, noviembre y diciembre se aproximan bastante.

4. Se obtuvo un modelo de series de tiempo que pasa todas las pruebas para la variable EMBUTIDOS:

$$\Delta\Delta_{12} y_t = (1 + 0{,}8436L^{12})(1 + 0{,}7492L)\xi_t$$

El pronósticos comparados con la venta real del los meses de octubre, noviembre y diciembre se aproximan bastante.

5. Se obtuvo un modelo de series de tiempo que pasa todas las pruebas pa variable CONGELADOS:

$$\Delta\Delta_{12} y_t = (1 + 0{,}9874L^{12})(1 + 0{,}6866L)\xi_t$$

El pronósticos comparados con la venta real del los meses de octubre, noviembre y diciembre se aproximan bastante.

6. Se obtuvo un modelo de series de tiempo que pasa todas las pruebas para la variable PRODUCTOS DEL MAR:

$$\Delta\Delta_{12} y_t = (1 + 0{,}9874L^{12})(1 + 0{,}6866L)\xi_t$$

El pronósticos comparados con la venta real del los meses de octubre, noviembre y diciembre se aproximan bastante.

7. Se obtuvo un modelo de series de tiempo que pasa todas las pruebas para la variable CONSERVAS:

$$\Delta y_t = (1 + 0{,}2236L + 0{,}2236L^2)\xi_t$$

El pronósticos comparados con la venta real del los meses de octubre, noviembre y diciembre se aproximan bastante.

8. Se obtuvo un modelo de series de tiempo que pasa todas las pruebas para la variable ARROZ:

$$\Delta\Delta_{12} y_t = (1 + 0{,}9623L^{12})(1 + 0{,}9464L)\xi_t$$

El pronósticos comparados con la venta real del los meses de octubre, noviembre y diciembre se aproximan bastante.

9. En el análisis de componentes principales 89% de la varianza total lo explica el primer componente, para la variable de CERDO. En este componente la variable más fuerte es FRESCURA.

10. En el análisis de componentes principales 88,2% de la varianza total lo explica lo explica el primer componente, para la variable de EMBUTIDOS. La variable SABOR es la predominante en esta línea.

11. En el análisis de componentes principales 78,3% de la varianza total lo explica los dos primeros componentes, para la variable de CONGELADOS. La variable de COSTUMBRE ingresa con fuerza en esta componente.

12. En el análisis de componentes principales 88,2% de la varianza total lo explica el primer componente, para la variable de PRODUCTOS DEL MAR. El EMPAQUE es una característica que influye en la compra.

13. En el análisis de componentes principales 90% de la varianza total lo explica el primer componente, para la variable CONSERVAS. Se identifica a la variable SABOR como la de mayor fuerza en el momento de la compra.

14. En el análisis de componentes principales 91,6% de la varianza total lo explica el primer componente, para la variable ARROZ. El SABOR y PRECIO son las variables que, según el estudio, influyen en la decisión de compra.

15. Para todas las variables: CERDO, EMBUTIDOS, CONGELADOS, PRODUCTOS DEL MAR, COMSERVAS Y ARROZ, las correlaciones entre las variables es muy notable, es decir todas las variables están correlacionadas.

16. Para la variable de CERDO, en la primera componente principal (que explica el 89,4% de la varianza) todas las variables originales tienen igual peso, por lo que a este eje lo denominaremos “Imagen de Marca”. A la segunda

componente principal la denominaremos “Eje del Precio”, puesto que la variable PRECIO es la que tiene mayor peso.

17. Para la variable de EMBUTIDOS, en la primera componente principa cual explica 88,2% de la varianza, todas las variables originales poseen pesos similares, por esta razón también denominaremos a este eje, “Imagen de Marca”. A la segunda componente principal, dado que la variable PRECIO vuelve a tener el mayor peso, se lo denominara “Eje del Precio”.

18. Para la variable CONGELADOS a la primera componente(que explica 88,8% de la varianza), también la denominaremos “Imagen de Marca” y a la segunda componente la denominaremos “Eje de la Textura” porque la variable original TEXTURA es la de mayor peso.

19. Para la variable PRODUCTOS DEL MAR en la primera componente principal, el cual explica 88,2% de la varianza, todas las variables originales poseen pesos similares, por esta razón también denominaremos a este eje, “Imagen de Marca”. A la segunda componente principal, dado que la variable CALIDAD tiene el mayor peso, se lo denominara “Eje de la Calidad”.

20. Para la variable CONSERVAS, en la primera componente principal, el cual explica 88,2% de la varianza, todas las variables originales poseen pesos similares, por esta razón también denominaremos a este eje, “Imagen de Marca”. A la segunda componente principal, dado que la variable CALI[illegible] tiene el mayor peso, se lo denominara “Eje de la Calidad”.

21. Para la variable ARROZ, en la primera componente principal, que explica 91,6% de la varianza, todas las variables originales poseen pesos similares, por esta razón denominaremos a este eje, “Imagen de Marca”. A la segunda componente principal, dado que la variable TEXTURA tiene el mayor peso, se lo denominara “Eje de la Textura”.

4.2. Recomendaciones

1. Las empresas de generalmente dirigen todas las campañas de mercadeo y ventas al segmento femenino de la población, este estudio demuestra que el sector masculino también es importante y se debería dirigirse a este segmento.

2. El análisis de componentes principales fue hecho en base a una muestra de la población de clientes atendidos por la empresa estudiada. Se recomienda utilizar el presente estudio, para realizar uno dirigido al consumidor final y obtener datos de la voz del consumidor.

APÉNDICES

APENDICE A

VENTA MENSUAL EN DOLARES POR LINEA DESDE MARZO DE 1994 A SEPTIEMBER 2006

Gráfico Apéndice A1

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Ventas Dólares Cerdo

	CERDO												
	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Enero		$ 95.355,80	$ 65.319,98	$ 33.348,61	$ 74.767,09	$ 70.647,96	$ 58.346,02	$ 79.099,94	$ 56.284,98	$ 111.636,41	$ 85.912,55	$ 114.596,22	$ 144.081,23
Febrero		$ 100.404,29	$ 68.514,59	$ 64.164,63	$ 77.820,74	$ 55.637,86	$ 106.658,36	$ 88.631,19	$ 102.065,52	$ 225.090,75	$ 100.340,44	$ 106.614,32	$ 109.144,82
Marzo	$ 98.851,75	$ 62.860,40	$ 86.919,57	$ 63.676,98	$ 68.416,83	$ 64.965,13	$ 87.830,08	$ 86.798,70	$ 102.150,10	$ 122.984,15	$ 120.352,08	$ 129.209,42	$ 105.944,65
Abril	$ 96.955,33	$ 107.220,89	$ 75.653,45	$ 64.143,74	$ 77.843,78	$ 71.740,94	$ 82.828,17	$ 97.421,43	$ 114.148,83	$ 113.985,33	$ 103.194,32	$ 115.950,75	$ 117.837,43
Mayo	$ 97.359,05	$ 139.186,17	$ 123.351,13	$ 52.965,44	$ 77.954,05	$ 82.177,26	$ 88.105,84	$ 115.346,63	$ 127.205,51	$ 194.233,15	$ 112.192,41	$ 129.526,54	$ 205.502,45
Junio	$ 96.087,43	$ 92.980,52	$ 78.158,26	$ 74.399,07	$ 80.002,86	$ 66.895,63	$ 83.834,20	$ 93.881,90	$ 109.522,58	$ 114.841,86	$ 108.412,55	$ 128.488,89	$ 115.694,36
Julio	$ 89.410,86	$ 101.547,35	$ 80.717,17	$ 64.950,61	$ 19.605,10	$ 57.439,62	$ 85.330,35	$ 92.535,53	$ 110.091,75	$ 82.402,06	$ 137.225,90	$ 119.423,88	$ 114.695,00
Agosto	$ 99.681,61	$ 100.757,22	$ 84.740,38	$ 64.932,58	$ 55.535,37	$ 73.306,73	$ 80.669,16	$ 95.311,60	$ 117.741,81	$ 102.394,35	$ 111.378,42	$ 138.844,71	$ 117.179,21
Septiembre	$ 101.437,20	$ 115.896,60	$ 88.818,35	$ 52.209,88	$ 78.764,81	$ 76.441,23	$ 75.604,83	$ 92.815,19	$ 116.235,98	$ 91.383,43	$ 101.300,47	$ 117.845,48	$ 115.564,23
Octubre	$ 80.779,06	$ 89.092,64	$ 78.439,73	$ 24.627,66	$ 88.453,17	$ 81.008,27	$ 80.074,35	$ 102.750,81	$ 122.058,36	$ 112.347,93	$ 119.241,18	$ 119.100,05	
Noviembre	$ 99.306,23	$ 141.969,89	$ 124.584,64	$ 72.608,79	$ 80.292,67	$ 82.999,03	$ 135.847,50	$ 95.186,79	$ 123.469,27	$ 124.755,04	$ 155.625,97	$ 140.689,52	
Diciembre	$ 105.264,60	$ 156.166,88	$ 130.813,87	$ 74.842,84	$ 86.716,09	$ 143.324,79	$ 146.715,30	$ 104.705,47	$ 185.109,68	$ 123.671,11	$ 224.350,44	$ 197.254,17	

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Gráfico Apéndice A2

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Ventas Dólares Embutidos

	EMBUTIDOS												
	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Enero		$ 21.396,63	$ 11.192,10	$ 6.566,45	$ 17.292,79	$ 25.298,02	$ 17.665,72	$ 21.040,45	$ 12.162,65	$ 22.907,82	$ 18.738,54	$ 22.714,53	$ 30.912,63
Febrero		$ 16.640,22	$ 20.788,95	$ 17.164,25	$ 16.214,21	$ 22.109,46	$ 32.627,90	$ 20.731,82	$ 21.086,83	$ 35.953,41	$ 23.806,34	$ 22.141,31	$ 21.034,24
Marzo	$ 25.844,08	$ 9.277,29	$ 21.122,21	$ 16.371,04	$ 12.995,94	$ 24.633,09	$ 22.393,59	$ 21.188,22	$ 20.466,22	$ 21.206,65	$ 24.613,71	$ 22.758,85	$ 21.433,37
Abril	$ 19.181,29	$ 15.319,17	$ 21.144,14	$ 16.788,81	$ 19.457,90	$ 25.982,78	$ 20.446,38	$ 20.540,88	$ 22.435,26	$ 24.408,21	$ 25.739,19	$ 22.652,19	$ 21.651,69
Mayo	$ 21.094,31	$ 18.649,07	$ 30.631,15	$ 10.429,07	$ 22.400,52	$ 27.615,43	$ 20.862,75	$ 29.134,69	$ 25.965,18	$ 40.129,72	$ 23.850,06	$ 23.629,08	$ 44.229,30
Junio	$ 21.056,08	$ 11.444,32	$ 20.315,69	$ 18.390,14	$ 24.556,31	$ 23.771,39	$ 20.682,27	$ 22.748,02	$ 24.789,22	$ 19.000,57	$ 24.146,42	$ 25.385,53	$ 23.482,20
Julio	$ 17.984,93	$ 12.297,33	$ 19.884,03	$ 17.803,96	$ 6.747,96	$ 21.058,22	$ 19.371,90	$ 19.752,05	$ 22.881,45	$ 19.598,97	$ 26.504,12	$ 25.173,20	$ 23.140,28
Agosto	$ 18.229,18	$ 12.354,05	$ 20.570,87	$ 19.939,28	$ 21.451,37	$ 26.511,99	$ 22.708,30	$ 19.949,57	$ 23.566,82	$ 25.413,06	$ 24.359,44	$ 27.278,51	$ 21.984,01
Septiembre	$ 18.303,76	$ 19.704,26	$ 21.393,04	$ 11.928,99	$ 24.750,02	$ 27.039,81	$ 21.276,38	$ 15.597,92	$ 22.894,65	$ 20.019,79	$ 21.262,46	$ 24.598,56	$ 21.565,56
Octubre	$ 15.057,12	$ 14.967,36	$ 22.490,18	$ 7.121,87	$ 26.851,78	$ 21.030,33	$ 20.775,41	$ 21.471,37	$ 22.305,11	$ 22.194,37	$ 24.324,27	$ 22.480,29	
Noviembre	$ 21.516,20	$ 19.022,05	$ 30.937,46	$ 20.233,53	$ 23.072,54	$ 27.891,58	$ 38.100,65	$ 22.042,84	$ 23.138,27	$ 26.678,12	$ 36.969,68	$ 24.682,01	
Diciembre	$ 22.807,17	$ 20.924,26	$ 32.484,33	$ 19.404,25	$ 24.918,34	$ 45.052,82	$ 41.148,70	$ 24.247,12	$ 37.452,10	$ 22.416,03	$ 41.414,60	$ 36.439,58	

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Gráfico Apéndice A3

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Ventas Dólares Congelados

	CONGELADOS												
	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Enero		$ 1.410,04	$ 1.024,48	$ 564,38	$ 2.169,30	$ 1.439,12	$ 2.068,30	$ 2.506,82	$ 1.493,44	$ 2.762,76	$ 2.306,67	$ 2.628,76	$ 4.041,68
Febrero		$ 1.275,54	$ 1.939,59	$ 1.855,28	$ 1.823,63	$ 1.416,58	$ 4.939,95	$ 3.219,29	$ 2.317,42	$ 3.943,71	$ 2.623,10	$ 1.968,82	$ 2.411,55
Marzo	$ 1.291,14	$ 1.276,62	$ 1.825,27	$ 1.551,73	$ 1.286,18	$ 1.798,49	$ 2.771,95	$ 2.581,92	$ 2.508,18	$ 2.113,18	$ 2.950,91	$ 2.510,38	$ 2.046,91
Abril	$ 1.343,93	$ 2.305,59	$ 2.077,88	$ 1.851,11	$ 2.153,35	$ 1.843,60	$ 2.632,03	$ 1.980,50	$ 2.411,38	$ 2.810,45	$ 3.583,76	$ 2.222,95	$ 2.353,24
Mayo	$ 1.616,09	$ 2.926,19	$ 2.408,63	$ 896,37	$ 2.128,39	$ 2.028,65	$ 2.514,36	$ 3.967,56	$ 2.693,40	$ 3.973,87	$ 2.926,03	$ 2.837,96	$ 4.455,33
Junio	$ 1.537,86	$ 2.155,49	$ 1.724,33	$ 1.999,75	$ 1.711,33	$ 1.644,34	$ 2.922,43	$ 2.669,78	$ 2.454,27	$ 1.399,01	$ 2.363,18	$ 2.056,62	$ 3.108,70
Julio	$ 1.518,36	$ 2.092,91	$ 1.444,21	$ 1.617,26	$ 470,60	$ 1.512,68	$ 3.214,95	$ 2.392,30	$ 2.649,65	$ 1.885,24	$ 3.917,74	$ 2.353,92	$ 1.974,77
Agosto	$ 1.445,91	$ 1.554,86	$ 1.808,08	$ 1.868,70	$ 1.564,62	$ 1.933,89	$ 2.667,56	$ 2.320,53	$ 2.725,55	$ 2.166,01	$ 3.353,29	$ 2.901,65	$ 2.293,94
Septiembre	$ 1.379,85	$ 2.486,89	$ 1.961,30	$ 870,36	$ 1.694,49	$ 2.171,33	$ 3.231,44	$ 1.842,81	$ 2.583,17	$ 2.086,66	$ 2.273,47	$ 2.386,17	$ 2.235,26
Octubre	$ 882,64	$ 1.599,17	$ 1.780,62	$ 801,15	$ 1.793,14	$ 1.951,24	$ 2.399,85	$ 2.493,14	$ 2.619,86	$ 2.384,50	$ 2.853,59	$ 2.401,78	
Noviembre	$ 1.648,41	$ 2.984,71	$ 2.432,72	$ 1.952,58	$ 2.192,24	$ 2.048,94	$ 4.534,53	$ 2.232,95	$ 2.939,89	$ 2.139,83	$ 3.720,38	$ 2.318,13	
Diciembre	$ 1.747,32	$ 3.283,19	$ 2.554,35	$ 1.795,13	$ 2.367,62	$ 4.392,27	$ 4.897,29	$ 2.456,25	$ 3.490,46	$ 2.295,06	$ 4.950,66	$ 4.038,27	

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Gráfico Apéndice A4

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Ventas Dólares Productos del Mar

	PRODUCTOS DEL MAR												
	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Enero		$ 1.556,30	$ 3.019,65	$ 1.093,78	$ 4.352,90	$ 3.234,01	$ 3.274,14	$ 4.013,48	$ 3.426,06	$ 6.188,97	$ 8.226,78	$ 6.380,51	$ 8.726,33
Febrero		$ 1.384,51	$ 2.428,69	$ 2.978,49	$ 3.843,32	$ 3.605,32	$ 5.657,11	$ 4.554,71	$ 5.241,80	$ 9.212,69	$ 7.803,21	$ 6.619,88	$ 5.766,84
Marzo	$ -	$ 2.519,52	$ 3.818,81	$ 3.844,65	$ 2.915,53	$ 3.548,03	$ 5.567,48	$ 4.321,28	$ 5.662,67	$ 5.094,81	$ 6.409,58	$ 6.266,89	$ 5.974,79
Abril	$ -	$ 3.626,91	$ 4.557,65	$ 2.856,37	$ 4.564,08	$ 3.839,71	$ 4.595,85	$ 4.290,67	$ 5.167,26	$ 5.671,31	$ 7.305,97	$ 5.286,38	$ 5.262,73
Mayo	$ 9,27	$ 3.242,08	$ 6.617,03	$ 1.737,18	$ 3.687,12	$ 4.011,25	$ 4.499,60	$ 7.337,77	$ 6.467,97	$ 10.051,00	$ 6.243,26	$ 6.411,99	$ 14.054,80
Junio	$ 448,94	$ 1.036,89	$ 4.680,12	$ 3.736,17	$ 3.316,71	$ 5.700,44	$ 3.826,43	$ 5.204,91	$ 4.991,62	$ 4.032,77	$ 7.479,92	$ 5.839,71	$ 6.767,29
Julio	$ 1.317,87	$ 1.418,18	$ 3.680,95	$ 2.940,66	$ 574,23	$ 4.783,32	$ 4.044,84	$ 4.737,59	$ 5.514,57	$ 5.259,49	$ 8.568,56	$ 6.482,55	$ 6.127,49
Agosto	$ 1.600,04	$ 1.170,79	$ 5.020,31	$ 3.710,44	$ 2.416,68	$ 5.780,11	$ 4.154,42	$ 4.690,21	$ 5.447,98	$ 5.703,49	$ 7.999,20	$ 7.470,41	$ 7.091,79
Septiembre	$ 1.074,28	$ 1.834,38	$ 4.080,67	$ 1.861,64	$ 3.204,66	$ 5.294,15	$ 4.755,37	$ 3.696,84	$ 6.255,37	$ 4.355,47	$ 5.996,44	$ 6.525,61	$ 6.743,31
Octubre	$ 1.002,72	$ 2.306,20	$ 3.965,53	$ 1.464,61	$ 3.146,49	$ 5.423,89	$ 3.991,96	$ 5.800,76	$ 5.755,53	$ 4.814,51	$ 5.913,12	$ 6.680,05	
Noviembre	$ 1.052,86	$ 3.306,92	$ 6.683,20	$ 3.232,55	$ 3.797,73	$ 4.051,36	$ 8.637,39	$ 5.432,10	$ 6.935,65	$ 4.795,18	$ 13.003,14	$ 7.089,18	
Diciembre	$ 1.760,04	$ 3.637,61	$ 7.017,36	$ 3.982,15	$ 4.101,55	$ 11.471,37	$ 9.328,38	$ 5.975,31	$ 10.554,51	$ 5.081,91	$ 10.736,81	$ 10.741,47	

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Gráfico Apéndice A5

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Ventas Dólares Conservas

	CONSERVAS												
	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Enero		$ 42.685,00	$ 28.969,08	$ 37.663,14	$ 93.274,16	$ 122.418,74	$ 83.203,41	$ 63.765,95	$ 47.740,49	$ 161.525,26	$ 93.117,69	$ 156.975,52	$ 198.051,61
Febrero		$ 53.085,57	$ 93.013,14	$ 74.055,19	$ 89.097,45	$ 77.896,67	$ 84.579,99	$ 98.125,83	$ 96.471,10	$ 196.505,79	$ 112.371,00	$ 154.994,50	$ 113.106,48
Marzo	$ 29.186,00	$ 50.584,00	$ 107.988,03	$ 72.396,40	$ 72.858,80	$ 56.399,46	$ 82.597,85	$ 96.007,80	$ 116.632,56	$ 177.674,43	$ 164.955,62	$ 104.562,95	$ 120.267,57
Abril	$ 54.445,26	$ 59.467,56	$ 66.176,76	$ 71.903,03	$ 85.212,63	$ 70.482,12	$ 65.551,67	$ 95.092,62	$ 117.539,92	$ 96.235,91	$ 158.168,28	$ 106.969,83	$ 118.694,87
Mayo	$ 42.030,06	$ 67.833,12	$ 101.816,24	$ 59.817,93	$ 73.267,44	$ 94.356,66	$ 82.318,57	$ 126.458,75	$ 102.709,61	$ 215.711,69	$ 124.514,25	$ 117.486,98	$ 219.294,05
Junio	$ 58.585,89	$ 51.226,40	$ 49.020,52	$ 72.569,55	$ 94.712,88	$ 34.686,27	$ 111.851,58	$ 143.631,14	$ 143.135,56	$ 87.138,56	$ 106.010,95	$ 144.270,10	$ 135.426,04
Julio	$ 19.400,40	$ 172.766,18	$ 73.814,35	$ 102.121,32	$ 38.989,94	$ 57.369,39	$ 67.584,93	$ 106.644,95	$ 113.115,74	$ 72.561,16	$ 80.634,99	$ 86.107,41	$ 125.058,37
Agosto	$ 49.462,92	$ 124.662,80	$ 89.807,01	$ 82.218,47	$ 60.940,41	$ 76.908,96	$ 117.660,80	$ 109.844,30	$ 102.165,50	$ 49.781,09	$ 114.408,55	$ 148.309,38	$ 108.838,47
Septiembre	$ 64.963,03	$ 57.495,23	$ 115.194,00	$ 73.452,17	$ 92.899,54	$ 138.043,38	$ 139.002,48	$ 106.652,55	$ 115.947,36	$ 96.121,40	$ 73.186,23	$ 142.724,38	$ 126.747,06
Octubre	$ 55.831,03	$ 75.793,11	$ 180.196,60	$ 22.265,90	$ 138.859,77	$ 89.533,67	$ 109.675,73	$ 103.475,98	$ 140.105,91	$ 88.897,93	$ 107.862,86	$ 146.794,82	
Noviembre	$ 42.870,66	$ 69.189,78	$ 102.834,40	$ 88.033,39	$ 75.465,46	$ 95.300,23	$ 213.524,95	$ 89.089,94	$ 125.662,05	$ 151.742,55	$ 175.819,46	$ 140.269,01	
Diciembre	$ 45.442,90	$ 76.108,76	$ 107.976,12	$ 75.584,89	$ 81.502,70	$ 141.392,49	$ 230.606,95	$ 97.998,93	$ 173.350,49	$ 218.023,68	$ 175.275,98	$ 187.577,73	

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

Gráfico Apéndice A6

"Análisis Multivariado de las Ventas de una Empresa de Alimentos Marzo 1994 – Septiembre 2006"

Ventas Dólares Arroz

	ARROZ												
	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Enero		$ 155.221,06	$ 137.410,15	$ 3.546,69	$ 32.293,19	$ 9.323,78	$ 11.242,06	$ 11.325,37	$ 5.684,55	$ 22.231,99	$ 13.158,43	$ 18.389,47	$ 67.751,57
Febrero		$ 127.218,73	$ 6.626,25	$ 49.191,42	$ 56.935,03	$ 8.852,87	$ 47.674,09	$ 14.328,54	$ 9.162,54	$ 33.347,68	$ 26.112,58	$ 19.968,87	$ 34.196,81
Marzo	$ 149,70	$ 64.502,53	$ 24.908,34	$ 37.995,63	$ 17.133,99	$ 9.997,13	$ 17.575,70	$ 13.389,85	$ 14.093,27	$ 16.678,18	$ 25.319,65	$ 20.233,47	$ 22.902,81
Abril	$ 154.291,91	$ 78.038,20	$ 17.314,78	$ 16.387,00	$ 31.751,65	$ 13.254,49	$ 13.197,32	$ 18.062,16	$ 15.180,08	$ 30.359,44	$ 21.795,13	$ 45.287,43	$ 19.948,22
Mayo	$ 78.192,85	$ 110.964,16	$ 22.141,19	$ 5.632,98	$ 9.191,15	$ 51.896,78	$ 15.931,87	$ 34.906,78	$ 15.934,67	$ 30.447,87	$ 13.362,25	$ 13.822,78	$ 48.518,50
Junio	$ 230.844,53	$ 52.968,85	$ 83.755,71	$ 49.211,06	$ 9.377,39	$ 15.939,77	$ 34.610,15	$ 16.346,63	$ 24.467,01	$ 26.213,82	$ 31.249,29	$ 22.723,84	$ 48.278,48
Julio	$ 54.459,08	$ 162.044,06	$ 10.638,06	$ 9.805,86	$ 44.381,82	$ 29.112,50	$ 29.790,73	$ 36.291,63	$ 19.697,81	$ 11.564,44	$ 27.213,07	$ 13.923,13	$ 17.566,64
Agosto	$ 103.783,36	$ 244.220,12	$ 29.817,13	$ 30.731,57	$ 7.562,84	$ 17.270,24	$ 59.011,01	$ 37.380,38	$ 18.707,03	$ 10.721,89	$ 23.239,21	$ 38.097,09	$ 17.005,46
Septiembre	$ 150.405,15	$ 16.764,19	$ 34.553,93	$ 18.169,00	$ 14.341,24	$ 41.595,23	$ 28.140,13	$ 18.489,95	$ 21.495,91	$ 17.201,53	$ 16.650,79	$ 22.585,16	$ 32.097,72
Octubre	$ 244.253,49	$ 14.524,69	$ 41.532,86	$ 3.053,47	$ 37.649,18	$ 13.425,20	$ 52.729,64	$ 12.620,51	$ 27.618,67	$ 19.807,39	$ 24.585,91	$ 31.555,44	
Noviembre	$ 79.756,71	$ 113.183,44	$ 22.362,60	$ 32.817,46	$ 9.466,88	$ 52.415,75	$ 28.720,10	$ 10.972,38	$ 10.270,94	$ 21.395,16	$ 46.884,44	$ 26.324,07	
Diciembre	$ 84.542,11	$ 124.501,79	$ 23.480,73	$ 38.587,74	$ 10.224,24	$ 29.948,97	$ 31.017,71	$ 12.069,62	$ 23.734,19	$ 12.966,14	$ 50.917,46	$ 44.732,14	

Autor: José Aguayo E. Fuente: Ventas de una Empresa de Alimentos Mar94-Sept06

BIBLIOGRAFIA

1. Carrasco Arroyo Salvador, Capítulo 2 Análisis de Componentes Principales, Estadística Multivariada, Universidad de Valencia – España.

2. Demetra Versión 2.0, desarrollado para Eurostat por Jens Dosse y Servais Hoffmann

3. Novales A., Econometría 2da. Mc. Grawhill 2000

4. Rencher A., 1998 “Maultivariate Statistical Analysis and Applications”, Wiley Series in Probability ans Statistics, New York.