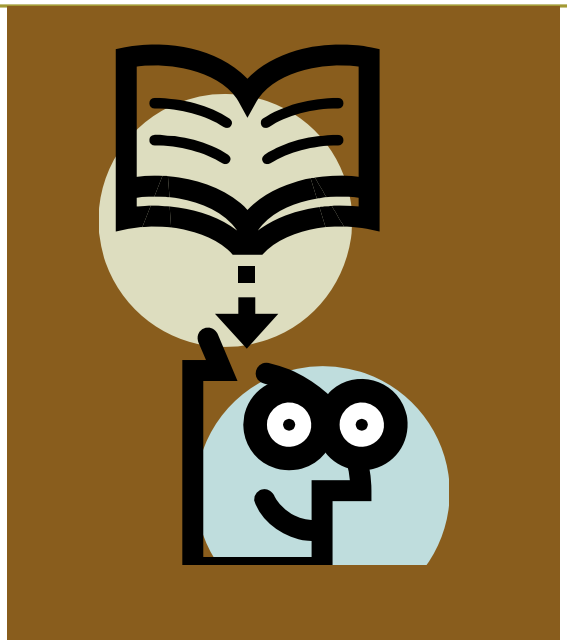


Análisis de la Información

PIE 5

ALFREDO MARIO BARONIO
ANA MARIA VIANCO

Cuadernos de Econometría

Análisis de la Información - PIE 5. Cuadernos de Econometría. Alfredo Mario Baronio y Ana María Vianco. 1ª edición. 2015

Xxx p; xxx cm

ISBN

1. *Regresión, 2 Factorial, 3 Independencia, 4 Pronóstico*

Fecha de catalogación: xxx 2015

Análisis de la Información - PIE 5. Cuadernos de Econometría.
Alfredo Mario Baronio y Ana María Vianco.

Primera edición xxx 2015

ISBN XXXXXXX

Tirada xxx ejemplares

Esta edición es financiada con subsidios otorgados al proyecto Producción de Datos y Econometría Aplicada por la Secretaría de Ciencia y Técnica de la UNRC y el Instituto de Investigaciones de la UNVM.

Contenido

En la última etapa del PIE –Proceso de Investigación Econométrica–, el investigador realiza el análisis de la información provista por la tabla de datos; que fue planteada en la segunda etapa y completada de acuerdo a lo establecido en la tercera y cuarta etapa.

En este cuaderno, dando respuesta a los objetivos planteados en la segunda etapa, se ven las técnicas apropiadas al análisis de la asociación entre las variables y se estudian los métodos para determinar la semejanza entre las unidades de observación. En ambos casos, los métodos para realizar estos estudios difieren de acuerdo a si, las mediciones sobre las unidades de observación, se han realizado con variables cualitativas o con variables cuantitativas o con ambas.

1. INTRODUCCIÓN AL ANÁLISIS DE INFORMACIÓN

El análisis de información debe ser previsto al principio del proceso de investigación econométrica. Ya sea que se trate de una encuesta o de información secundaria, existen varios tipos de análisis de la información que se pueden realizar con la tabla de datos completa. Esto facilita la evidencia empírica que desea realizar el investigador. Para un proceso completo de evidencia se deben aplicar análisis exploratorios, descriptivos, explicativos y confirmatorios.

1.1 Análisis exploratorio y descriptivo

Estos análisis, se pueden realizar de forma univariada, bivariada o multivariada. Esto es, dada una tabla de datos, se puede:

- Analizar cada variable de forma aislada o estudiar su relación (o asociación) con otra u otras variables. Cada uno de estos análisis es diferente si las unidades de observación son de corte transversal, de tiempo o de panel; de igual modo difiere si se trata de variables cuantitativas o cualitativas.
- Estudiar la semejanza entre las unidades de observación, que también dependerá si se trata de variables cuantitativas o cualitativas.

En este cuaderno se presentarán los análisis univariado y bivariados, tanto para variables observadas en unidades de corte transversal como

de tiempo. Se estudiarán, además, los análisis multivariados para unidades de observación.

Análisis univariados

Ya sea que se haya originado en una encuesta o en una fuente de información secundaria, cada variable estará asociada a una unidad de observación que resultará útil para realizar análisis univariados que incluyen varios aspectos:

- Representación (tablas/gráficos) de la serie observada sobre el conjunto de las unidades de observación y construcción eventual de una distribución de frecuencias.
- Resúmenes numéricos para las variables cuantitativas (valor central, medida de dispersión y de asimetría) y construcción de box plot.
- Examen de valores extremos y de anomalías aparentes.
- De tratarse de unidades de observación temporales: orden de integración de la variable analizada y análisis de tendencias.
- Primeras interpretaciones de esos análisis.

Análisis bivariados

Los vínculos entre dos variables deben ser seleccionados y estudiados adecuadamente. Se aconseja determinar cuanto antes los vínculos deseados por el investigador. Comentan Droesbeke y Fine (1997), que "el método que consiste en cruzar todas las parejas de variables no sólo es costoso sino además completamente innecesario".

El análisis consiste en determinar medidas de asociación en función del tipo de variables utilizadas:

- Coeficiente de correlación entre dos variables cuantitativas
- Razón de correlación entre variable cuantitativa y variable cualitativa
- Coeficiente χ^2 o tablas de contingencia para variables cualitativas

Análisis multivariados

La posibilidad de recurrir a métodos exploratorios y descriptivos multidimensionales puede determinar estructuras interesantes entre las unidades de observación. Los métodos más utilizados son:

El análisis factorial de correspondencias múltiples
El análisis en componentes principales
Los métodos de clasificación y agrupamiento de unidades de observación

1.2 Análisis explicativo

La relación de causalidad entre las variables puede implicar la aplicación de regresiones simples o regresiones múltiples, lineales o no lineales; en este último caso se estudia el efecto o el impacto de la variación de dos o más variables sobre la variación de otra variable (análisis de regresión múltiple). Eventualmente, se pueden especificar y estimar modelos multiecuacionales.

Este tipo de análisis es la base de la Econometría tradicional y de la nueva Econometría. Las técnicas estadísticas de regresión simple y múltiple son el fundamento con el que se desarrolló la econometría tradicional basada en el análisis y comprobación de modelos económicos, previamente especificados por el investigador. La nueva Econometría, a través del Proceso Generador de Datos (PGD) vincula los análisis descriptivos y exploratorios a la comprobación de relaciones entre variables económicas, en espacio y tiempo determinado.

Aunque no será tema central de este cuaderno, se puede decir que el PGD basa el análisis explicativo en cinco etapas. En cada una de ellas aplica técnicas estadísticas univariadas, bivariadas y multivariadas.

1.3 Análisis confirmatorio

Se pueden realizar *análisis confirmatorios* utilizando procedimientos basados en la estadística inferencial para estimar los parámetros poblacionales de las variables analizadas o de los modelos especificados, a partir de la información de la muestra o de fuentes secundarias descrita en la tabla de datos. Los métodos que se utilizan son:

- Estimación puntual y por intervalos de confianza
- Test de hipótesis relativo a parámetros de una población o de varias
- Estimación y test de hipótesis relativos a los parámetros de un modelo

La aplicación de estos métodos depende de numerosos factores, entre ellos, del tipo de muestreo utilizado (¿es la muestra aleatoria simple?), de la verificación de las hipótesis subyacentes a la aplicación de un test, de si el modelo especificado cumple con los supuestos sobre la perturbación aleatoria, etc. Cuestiones por las que el investigador debe ocuparse y verificar.

En principio y de acuerdo a proceso dispuesto en este libro (PIE), lo primero que debe hacer el investigador es confirmar las hipótesis estadísticas y económicas que llevaron a la realización de su estudio. Para ello, la inferencia estadística provee técnicas que son de utilidad para realizar este análisis. Las mismas se basan en análisis univariado e incluyen la aplicación de los conceptos de estimación puntual y por intervalo, de error de muestreo y tipo de muestreo aplicado en el trabajo de investigación.

A continuación se incluyen ejemplos de cuándo y cómo aplicar cada técnica estudiada en el cuaderno N° 3. Dada su tabla de datos, el investigador debe responder a las siguientes preguntas para

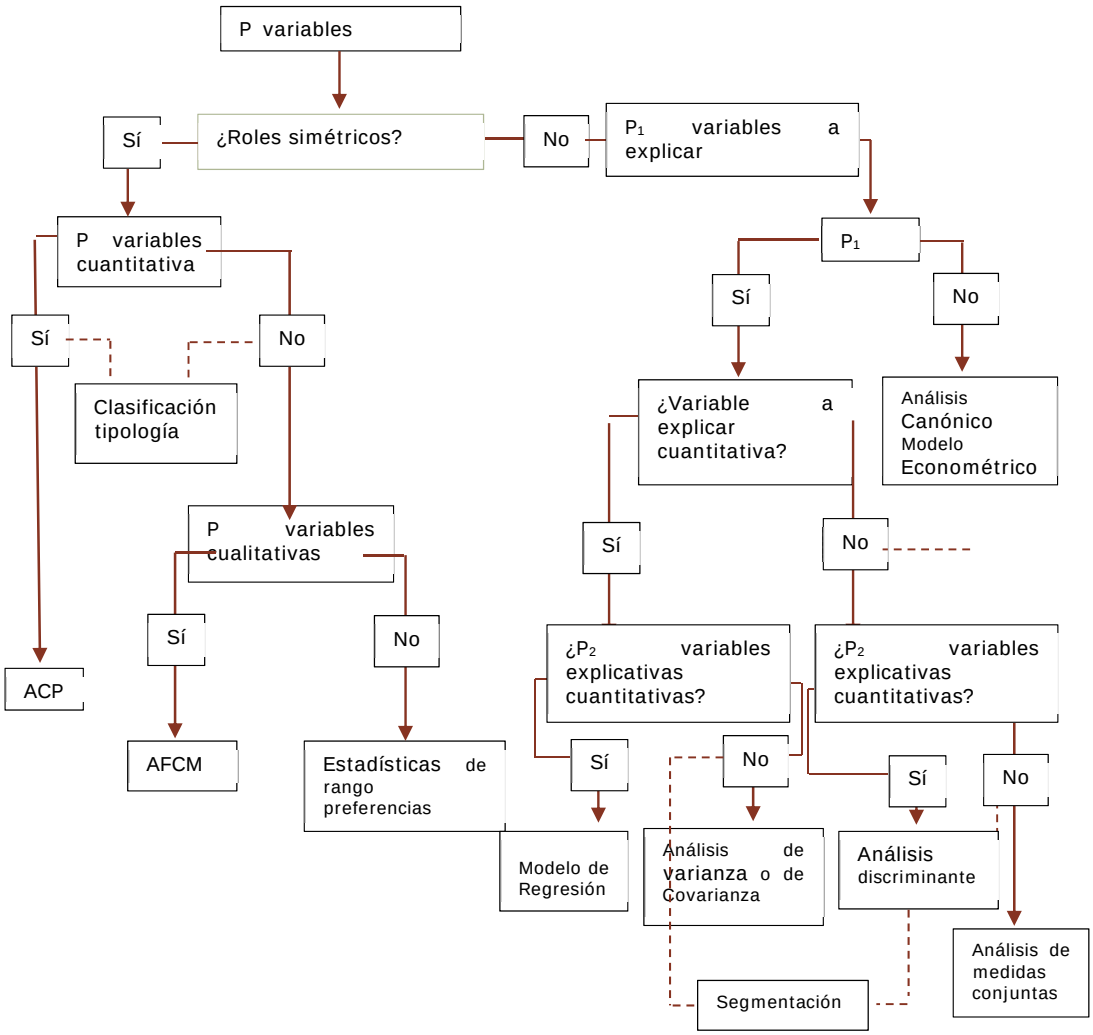
“confirmar” el resultado de la investigación de acuerdo a sus hipótesis iniciales:

a) ¿Qué tipo de muestreo utilizó?
b) ¿Qué parámetro poblacional está estimando?
c) ¿Cuál es el error de muestreo que cometió en la estimación?
d) ¿Cuál es el intervalo de confianza para el parámetro estimado en la investigación?
e) ¿Cumple ese parámetro con las hipótesis estadísticas y económicas formuladas al inicio de la investigación?
f) ¿Cuál es el efecto de diseño que obtuvo con esa estimación?

1.4 Panorama actual de los métodos de análisis de la información

Muchas veces el investigador duda sobre el comportamiento que debe adoptar frente a la elección de un método de tratamiento estadístico. Suele disponer de manuales y paquetes estadísticos que presentan una lista más o menos impresionante de técnicas frente a la cual queda sorprendido. Sin querer ser exhaustivo, se puede aclarar la elección consultando “mapas” de métodos como el presentado a continuación.

La lectura es simple. Pero “práctica y experiencia” son las únicas maneras de manejarlos correctamente. Se ha elegido deliberadamente presentar técnicas que no toman en cuenta una dimensión temporal. El lector interesado en este último caso podrá consultar, por ejemplo, Coutrot y Droesbeke (1995) y Menard (1991).



OBSERVACIÓN. Siguiendo a Droesbeke y Fine (1997), si se hubiera de resumir en unas reglas elementales el comportamiento del investigador, se podría proponer lo siguiente:

Regla 1: De nada sirve analizar un problema que no se entiende. Intente también saber de dónde vienen los datos a tratar y porqué han sido elegidos. Busque las informaciones complementarias susceptibles de mejorar su comprensión.

Regla 2: Examine la manera en que se colectaron los datos: el método de recogida tiene una influencia no despreciable sobre el tratamiento de los datos.

Regla 3: Examine cuidadosamente la estructura de los datos: ¿De qué variables se trata? ¿Miden una realidad o sirven de indicador de una situación? ¿Desempeñan un papel simétrico o no? ¿Se trata de variables a interpretar o de variables de control? ¿Qué escalas de medidas se utilizaron? ...

Regla 4: Nunca hay que economizar un análisis exploratorio en el que se utilizan los medios más elementales para ver y sentir los datos. Piense en particular la utilización de (buenos) gráficos para ayudarle en esta vía.

Regla 5: Desconfíe de los métodos de tratamiento "recetas". Recurrir a paquetes estadísticos impide a menudo examinar la pertinencia de los métodos utilizados.

Regla 6: Desde el principio del análisis hasta el informe final utilice la técnica del va y viene, las vueltas atrás suelen permitir evaluar la coherencia. El informe final debe ser claro, fácilmente legible y preciso.

Regla 7: Utilice constantemente un instrumento del cual no se ha hablado ¡el sentido común!

2. ANÁLISIS UNIVARIADO PARA SERIES ESTADÍSTICAS

2.1 Componentes de una serie

En buen número de circunstancias, a la hora de describir o predecir un fenómeno social o económico, se cuenta con información histórica organizada en series estadísticas denominadas series temporales (serie cronológica o histórica). Así se denomina a las variables observadas para un número determinado de unidades de observación temporales. Es decir, una serie temporal es un conjunto de datos sobre una variable determinada, para distintos momentos de tiempo.

Ejemplo 2.1. Los datos anuales de la Producción de Cereales, los trimestrales del PIB, los indicadores mensuales de coyuntura económica, la evolución de ventas de empresa, etc.

Un concepto relacionado con los datos de series temporales es el de *frecuencia*: período de tiempo que separa dos de sus datos. Cuanto menor es el período transcurrido, mayor es la frecuencia de los datos. En general, la unidad de observación temporal es de frecuencia regular y hace referencia al momento de tiempo anual, mensual o trimestral.

La elección de una u otra técnica de análisis o predicción depende del supuesto establecido sobre el esquema de descomposición y la identificación de los componentes de una serie.

Una serie estadística se puede descomponer en cuatro componentes: tendencia (T), ciclo (C), estacionalidad (S) y componente irregular (I), atendiendo a *dos esquemas de descomposición*.

- Según el *esquema multiplicativo*, la serie económica es el resultado del producto de sus componentes, esto es:

$$Y = T * C * S * I$$

- Según el *esquema aditivo*, la serie es el resultado de la suma de sus componentes:

$$Y = T + C + S + I$$

En la práctica, se suele considerar que los cuatro componentes teóricos quedan reducidos a dos: uno que recogería movimientos de la serie a largo plazo – tendencia o tendencia más ciclo – y otro que reflejaría variaciones de la serie por razones de estacionalidad. Los movimientos cíclicos raramente tienen un tratamiento independiente y los de carácter irregular forman el residuo o término de error que servirá para analizar la propia bondad de la aplicación realizada. En definitiva, la serie se considera formada por:

$$Y = T + S$$

Significa que el valor original de una serie es el resultado de multiplicar un valor de tendencia por un factor de estacionalidad (de media 1 o 100, en índice). No obstante, los tratamientos que se muestran en este cuaderno son generalizables al esquema aditivo, o incluso a esquemas mixtos.

Tendencia (T)

La tendencia es aquel componente de una serie estadística vinculada al movimiento a largo plazo de la misma. Se trata del patrón regular de comportamiento a largo plazo, sea éste creciente o decreciente.

Ejemplo 2.2. La inflación es un proceso inherente a casi todas las economías en mayor o menor grado. Por ello, cabe esperar que el patrón de comportamiento regular de los índices de precios sea

creciente. Se espera a largo plazo que los precios suban. Por ello, la serie estadística que recoge la evolución de los precios debe presentar un importante componente tendencial. Así aparece reflejado en el gráfico de la Figura 2.1, donde se recoge la evolución del IPC argentino (nivel general) desde 2004 hasta 2007 (promedio anual), base 1999=100.

La clave es comprender que, más allá de subas y bajas parciales (de un mes a otro) existe una clara tendencia al crecimiento de esta variable.

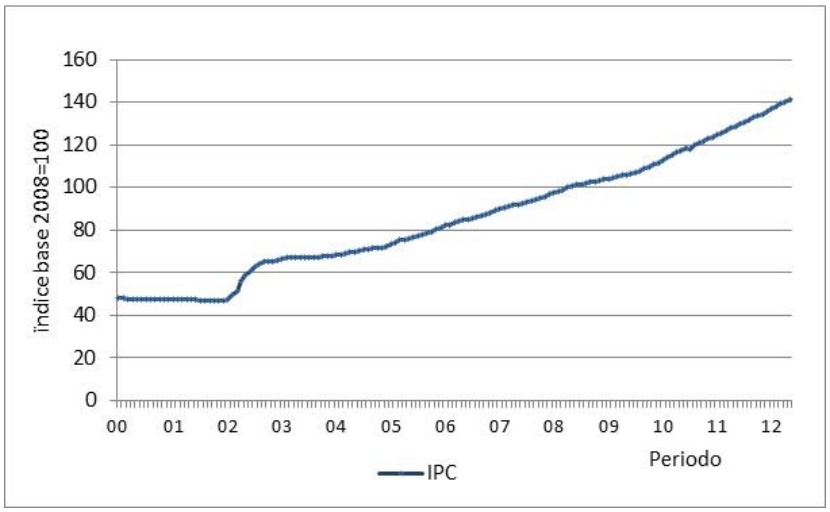


Figura 2.1 Índice de Precios al Consumidor

Ciclo (C)

El ciclo es aquel componente de la serie vinculado a oscilaciones a medio plazo. Se suele considerar que no presenta un movimiento con patrón uniforme único, como la tendencia o la estacionalidad, por lo que son oscilaciones en períodos superiores al año; este aspecto sirve, entre otros, para distinguirlo de la estacionalidad. En la práctica, se supone que en una serie económica se solapan distintos ciclos; por esto, es habitual considerar conjuntamente el ciclo con la tendencia, en un único movimiento de la serie a largo plazo, frente a movimientos causados por la estacionalidad.

Estacionalidad (S)

La estacionalidad es una componente que se presenta en series de frecuencia inferior a la anual (mensual, trimestral,...), y supone oscilaciones a corto plazo de período regular, inferior al año y amplitud regular. Se trata de la componente que introduce elementos interesantes para la predicción. En general, las series de frecuencia inferior a la anual presentan estacionalidad.

Ejemplo 2.3. En el gráfico de la Figura 2.2 se recoge la evolución del Índice de Producción industrial de Río Cuarto y Argentina, desde enero de 2005 hasta febrero de 2008. Son datos mensuales, en base 2004=100, obtenidos del Estimador Mensual Industrial que elabora el INDEC para Argentina y la Municipalidad de Río Cuarto para la ciudad. Se puede discutir si la serie presenta tendencia o ciclo, realmente es un período muy pequeño para detectarlo gráficamente; pero obsérvese que la serie presenta un comportamiento en los distintos meses de cada año muy similar. Todos los años la serie parece crecer en los primeros meses, decae ligeramente en abril, y, sobre todo, sufre una drástica caída de forma sistemática en el mes de enero. Y así año tras año. Se detecta pues un patrón regular. Efectivamente, esa evolución de la serie responde a hechos sencillos. En los meses de enero y de abril (si en este mes cae la Semana Santa) las empresas registran la disminución en su producción debido a las vacaciones. Eso es estacionalidad.

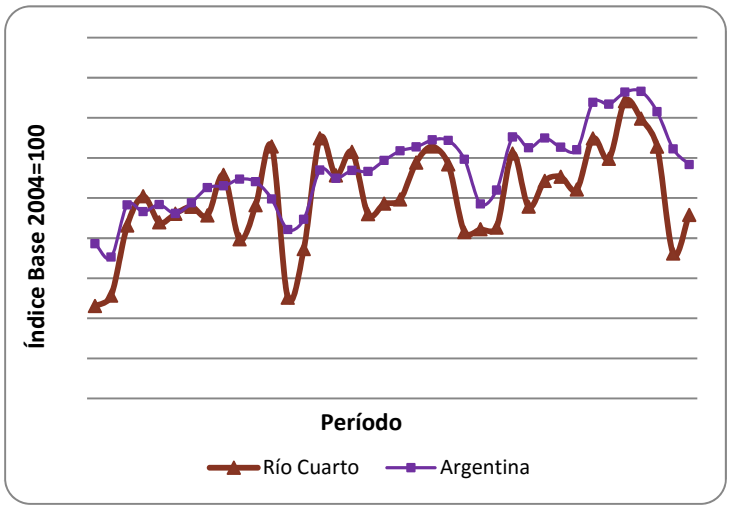


Figura 2.2 Estimador Mensual Industrial (EMI)

Los procedimientos para obtener el componente estacional de la serie se analizarán con detalle posteriormente.

Componente irregular (I)

También llamado residual o aleatorio. El aspecto importante de este componente es que no sigue algún patrón sistemático de comportamiento que se pueda modelizar. En todo caso, los residuos se utilizan para comprobar la bondad de la aplicación realizada.

El gráfico de la Figura 2.3 ilustra los componentes de la serie de Inversión interna bruta fija en equipos durables, en millones de pesos a precios de comprador de 1993, entre el primer trimestre de 1993 y el cuarto trimestre de 2012.

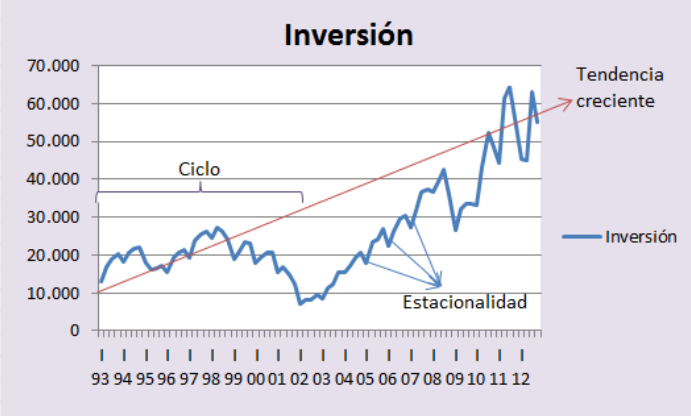


Figura 2.3

A continuación se estudiarán los métodos de tendencia, suavización y pronóstico de series estadísticas temporales:

- mínimos cuadrados ordinarios
- medias móviles
- alisado exponencial sin tendencia
- alisados con tendencia

2.2 Introducción a los métodos de estudio de una serie de tiempo

El factor componente de una serie de tiempo que se estudia con más frecuencia es la tendencia porque resulta útil a la hora de realizar pronósticos.

Para obtener alguna impresión visual de los movimientos generales a largo plazo en una serie de tiempo, se construye una gráfica en la cual se trazan los datos observados para la variable (variables dependientes) y los períodos de tiempo (unidades de observación, consideradas en este análisis como si fueran variables independientes).

Si fuera posible ajustar en forma adecuada una *tendencia lineal* a los datos, los dos métodos en uso más extenso para ajustes de tendencias, son el de los mínimos cuadrados y el de la doble suavización exponencial.

Sin embargo, si los datos de la serie de tiempo indican algún movimiento curvilíneo descendente o ascendente a largo plazo, los dos métodos en mayor uso para el ajuste de tendencias son el de los mínimos cuadrados y el de la triple suavización exponencial.

Mínimos Cuadrados

El método de los *mínimos cuadrados* permite ajustar una línea recta de la forma:

$$\hat{Y}_t = \hat{\alpha} + \hat{\beta} X_t$$

De modo que los valores que se calculen para los dos coeficientes cumplan con el siguiente requisito: minimizar la suma de las diferencias al cuadrado entre cada valor observado en los datos y cada valor estimado a lo largo de la línea de tendencia, es decir:

$$\sum_{t=1}^{T}\left(Y_t - \hat{Y}_t\right)^2 = mínimo$$

Para obtener la recta estimada se calcula la pendiente y la intercepción con:

$$\hat{\beta} = \frac{\sum_{t=1}^{T} X_t Y_t - \frac{\left(\sum_{t=1}^{T} X_t\right)\left(\sum_{t=1}^{T} Y_t\right)}{T}}{\sum_{t=1}^{T} X_t^2 - \frac{\left(\sum_{t=1}^{T} X_i\right)^2}{T}} = \frac{\sum_{t=1}^{T}(X_t - \bar{X})(Y_t - \bar{Y})}{\sum_{t=1}^{T}(X_t - \bar{X})^2}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

Una vez obtenida la línea $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$, se pueden substituir los valores de X_t, en este caso *tiempo*, para predecir diversos valores para Y_t. Al trabajar con datos de una serie de tiempo deben codificarse los valores de X_t asignándole códigos crecientes de enteros: 1, 2,…,t,..., T.

Aislamiento y eliminación de la tendencia

Se estudió la tendencia como una ayuda para el pronóstico a corto plazo. Pero los investigadores, también pueden desear el estudio de la tendencia de modo que, los factores que influyen en ella, se puedan eliminar del modelo multiplicativo de series de tiempo clásico y, por tanto, proveer la estructura para el pronóstico a corto plazo de la variable en cuestión. El procedimiento de aislar y eliminar un factor componente de los datos, se llama *descomposición de las series de tiempo*.

Como el método de los mínimos cuadrados provee valores $\hat{Y}_t$ de tendencia "ajustados" para cada año en la serie, se puede eliminar con facilidad la componente de tendencia del modelo multiplicativo de

series de tiempo clásico (porque en cualquier año dado la componente de la tendencia, se estima con $\hat{Y}_t$).

Por tanto, la componente de la tendencia se puede eliminar mediante una división en el modelo multiplicativo

$$\frac{Y_t}{\hat{Y}_t} = \frac{T_t\, C_t\, I_t}{\hat{Y}_t}$$

Pero como $\hat{Y}_t = T_t$ se tiene:

$$\frac{Y_t}{\hat{Y}_t} = \frac{T_t\, C_t\, I_t}{\hat{Y}_t} = C_t\, I_t$$

Relativas cíclicas e irregulares

Las proporciones entre los valores observados y los valores de las tendencias ajustadas, $Y_t/\hat{Y}_t$ que se calculan cada año en la serie, se llaman *relativas cíclicas-irregulares*. Estos valores, que fluctúan alrededor de 1 muestran la componente tanto cíclica como irregular en la serie.

Ejemplo 2.4. La serie de tiempo presentada en la tabla y trazada en la Figura 2.4, representa los pagos anuales (en miles de millones de dólares) a las compañías de seguros de vida, tanto por intereses sobre préstamos con garantía de la póliza como por primas fraccionadas en el período de 10 años de 1967 a 1976.

PRÉSTAMOS ANUALES SOBRE PÓLIZAS Y PRIMAS FRACCIONADAS DE COMPAÑÍAS DE SEGUROS DE VIDA (1967 - 1976)

Año t	X_t	Y_t	$X_t - \bar{X}$	$Y_t - \bar{Y}$	$(X_t - \bar{X})(Y_t - \bar{Y})$	$(X_t - \bar{X})^2$
1967	1	10,1	-4,5	-7,9	35,5	20,3
1968	2	11,3	-3,5	-6,7	23,4	12,3
1969	3	13,8	-2,5	-4,2	10,5	6,3
1970	4	16,1	-1,5	-1,9	2,8	2,3
1971	5	17,1	-0,5	-0,9	0,4	0,3
1972	6	18	0,5	0,0	0,0	0,3
1973	7	20,2	1,5	2,2	3,3	2,3
1974	8	22,9	2,5	4,9	12,3	6,3
1975	9	24,5	3,5	6,5	22,8	12,3
1976	10	25,9	4,5	7,9	35,6	20,3
T = 10	55	179,9	0	0,0	146,7	82,5

Se determina que:

$$\hat{\beta} = \frac{\sum_{t=1}^{T}(X_t - \bar{X})(Y_t - \bar{Y})}{\sum_{t=1}^{T}(X_t - \bar{X})^2} = \frac{146.7}{82.5} = 1.8$$

y dado que:

$$\bar{Y} = \frac{\sum_{t=1}^{T} Y_t}{T} = \frac{179.9}{10} = 17.99$$

$$\bar{X} = \frac{\sum_{t=1}^{T} X_t}{T} = \frac{55}{10} = 5.5$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 17.99 - 1.8 * 5.5 = 8.2$$

Se tiene:

$$\hat{Y}_t = 8.2 + 1.8\, X_t$$

La intercepción $\hat{\alpha} = 8.2$, es el valor ajustado de la tendencia que refleja la cantidad de dinero (en miles de millones de dólares) pagado a las compañías de seguros de vida por intereses por préstamos con garantía de la póliza y por primas fraccionadas durante el año 1966. La pendiente $\hat{\beta} = 1.8$, indica que tales pagos van aumentando a razón de 1.800 millones de dólares por año.

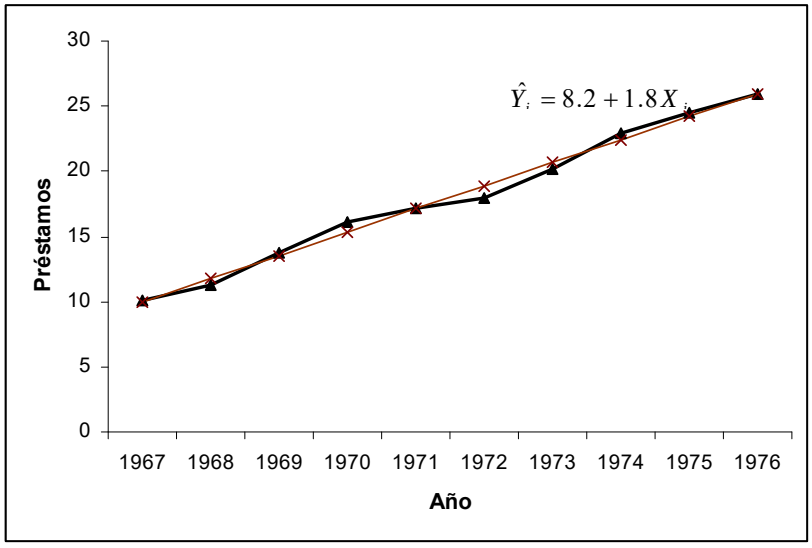


Figura 2.4. Pagos anuales a las compañías de seguro – Evolución y Tendencia

Para ajustar la línea de tendencia a los años observados de la serie, sólo se substituyen los valores codificados correspondientes a X_t en la ecuación. Por ejemplo, en el año 1975, en donde $X_{1975} = 9$, el valor predicho (ajustado) de la tendencia, se expresa con:

$$\hat{Y}_{t=1975} = 8.2 + 1.8 * 9 = 24.2 \; miles\ de\ millones\ de\ dólares$$

Para usar la recta de tendencia en predicción, se puede proyectar la línea ajustada mediante extrapolación matemática. Por ejemplo, para predecir la tendencia en los pagos para el año 1977, se substituye $X_{1977} = 11$, el código para el año 1977, en la ecuación y se pronostica que la tendencia será:

$$\hat{Y}_{t=1977} = 8.2 + 1.8 * 11 = 28 \; miles\ de\ millones\ de\ dólares$$

Los cálculos de las relativas cíclicas-irregulares se muestran en la siguiente tabla, para los datos de las series de tiempo de 10 años que reflejan los pagos anuales a las compañías de seguros por intereses de préstamos con garantía de póliza y por pago de primas fraccionadas. Los valores de tendencia ajustada [columna (4)], se determinan substituyendo los valores apropiados con el código X [columna (2)], en el modelo de tendencia lineal obtenido con el método de los mínimos cuadrados. Para cada año de la serie, el valor observado [columna (3)] se divide del valor de tendencia ajustada [columna (4)] para producir la relativa cíclica-irregular [columna (5)].

OBTENCIÓN DE LAS RELATIVAS CÍCLICAS – IRREGULARES

$Año\ t$ (1)	X_t (2)	Y_t (3)	$\hat{Y}_t = 8.2 + 1.8\ X_t$ (4)	$Y_t / \hat{Y}_t$ (5)
1967	1	10,1	10	1,010
1968	2	11,3	11,8	0,958
1969	3	13,8	13,6	1,015
1970	4	16,1	15,4	1,045
1971	5	17,1	17,2	0,994
1972	6	18	19	0,947
1973	7	20,2	20,8	0,971
1974	8	22,9	22,6	1,013
1975	9	24,5	24,4	1,004
1976	10	25,9	26,2	0,989

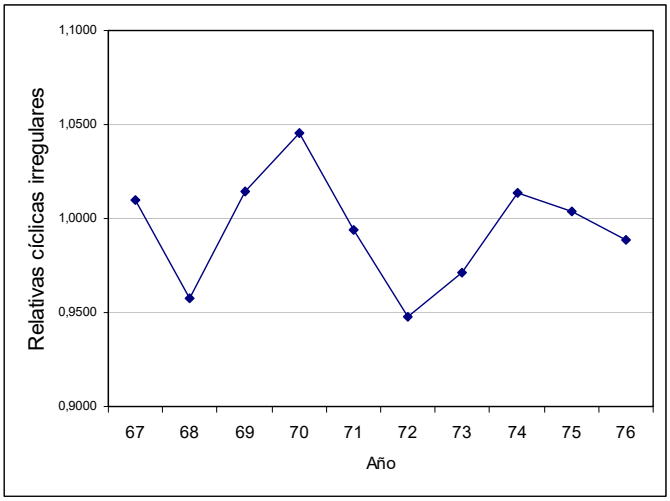


Figura 2.4. Pagos anuales a las compañías de seguro – Relativas cíclicas-irregulares

Medias móviles

La técnica de medias móviles construye una nueva serie a partir de la media de un número determinado de datos, en la que se va añadiendo sucesivamente un dato nuevo y quitando, al mismo tiempo, el más antiguo de los datos incluidos en la media anterior. Se trata de una técnica que utiliza información estadística pasada, y es considerada como técnica naïve (en sentido amplio) y de alisado.

OBSERVACIÓN.

Naïve

En general y aplicado a múltiples contextos, es frecuente denominar como *naïve* (ingenuo) a aquel procedimiento de predicción que repite de forma mecánica un comportamiento

pasado. De hecho, gran parte de las técnicas que se estudian aquí se pueden calificar de *naïve*.

Por ejemplo, el conocido como *modelo naïve I* realiza predicción asumiendo que el valor futuro (en el período t + 1) de una variable (es decir, el valor predicho, $\hat{y}_{t+1}$) coincide con el valor actual (el valor de la variable en el momento t), es decir:

$$\hat{y}_{t+1} = y_t$$

Una versión alternativa, conocido como *modelo naïve II*, asumiría no la igualdad del valor sino del incremento, es decir:

$$\hat{Y}_{t+1} - y_t = y_t - y_{t-1}$$

Obviamente, se trata de procedimientos muy simples, por lo que se puede utilizar como predicción el valor medio de un período, en lugar de sólo el último:

$$\hat{Y}_{t+1} = \bar{y}$$

Sin embargo esto sólo sería aconsejable cuando la variable o serie no tuviese tendencia, sino que oscilase aleatoriamente en torno a la media. *En ese caso, la media es el valor más probable de predicción.*

Como la mayor parte de las series económicas presentan tendencia (piénsese en una serie como el PIB, que más allá de caídas puntuales, muestra tendencia ascendente a lo largo del tiempo), se pueden elaborar alternativas igualmente "ingenuas". Una primera consiste en añadir en el modelo *naïve* I un término de tendencia, dada una tasa c de crecimiento constante, es decir:

$$\hat{Y}_{t+1} = (1 + c) \cdot y_t$$

Una segunda opción consiste en eliminar la tendencia de la serie, predecir la nueva serie y posteriormente añadir la tendencia.

Para ello suele ser suficiente con calcular la primera diferencia de la serie; es decir, si Δy_t representa la serie en diferencias, se trata de calcular para cada momento de tiempo la expresión:

$$\Delta y_t = y_t - y_{t-1}$$

En todo caso, si se trabaja con medias aritméticas como predictor aplicadas a la serie en incrementos, se estaría utilizando la media de los incrementos, que es igual al recorrido de la serie en niveles (primer valor menos último), dividido por el número de períodos. Es decir:

$$\overline{\Delta y}_t = \frac{[(y_t - y_{t-1}) + (y_{t-1} - y_{t-2}) + \cdots + (y_1 - y_0)]}{n} = \frac{(y_t - y_0)}{n}$$

Con lo que para predecir el incremento en la serie se haría:

$$\Delta \hat{y}_{t+1} = \overline{\Delta y}_t = \frac{(y_t - y_0)}{n}$$

De forma que la predicción de la serie en niveles adoptaría la expresión:

$$\hat{y}_{t+1} = y_t + \Delta \hat{y}_{t+1} = y_t + \frac{(y_t - y_0)}{n}$$

La práctica operativa de este procedimiento exigiría calcular, período a período, la media de los incrementos en los valores de la serie para todo el período muestral para actualizar la predicción. Ello introduce, además de falta de economicidad, dos problemas:

cuando el período muestral es muy extenso se corre el riesgo de que cambios estructurales (es decir, cambios sustanciales en el comportamiento del fenómeno) acontecidos en un momento determinado no queden reflejados correctamente, puesto que el valor de predicción se forma con todo el período muestral y no sólo con los más recientes.

con períodos muestrales cortos se presenta el problema de la estacionalidad. Algunas series económicas presentan valores atípicos (por exceso o por defecto) en determinados períodos del año. Por ejemplo, la producción industrial cae de forma muy relevante en enero. A la hora de predecir el valor de tal serie en un mes como febrero, la inclusión de un mes como enero en el cálculo de la media puede infravalorar el valor de predicción.

Por todo ello, es más frecuente utilizar la técnica conocida como *medias móviles*.

Alisado

Reciben este nombre aquellas técnicas, como las medias móviles o el alisado exponencial, que “alisan”, en el sentido de moderar, las variaciones que una serie económica pueda presentar, sean estas estacionales (sólo en determinados momentos del año), cíclicas (recurrentes cada ciertos años, es decir, debidas al momento del ciclo económico) o irregulares. Con frecuencia resulta de interés eliminar estos comportamientos, puesto que introducen ruido y, con frecuencia, no ayudan en la predicción. En última instancia, estas técnicas provocan que la serie presente un comportamiento más estable.

La técnica de medias móviles, en sus múltiples variantes, provoca esta circunstancia en las series económica, tanto más cuanto mayor sea el número de términos en la media móvil. Esto se observa con claridad con un ejemplo como el recogido en la Figura 2.5. Se ha representado la serie original utilizada en el Ejemplo 2.5 (a continuación), junto con la media móvil simple de tres (M3t) y de seis términos (M6t). Aprecie cómo, efectivamente, la técnica modera las variaciones de la serie.

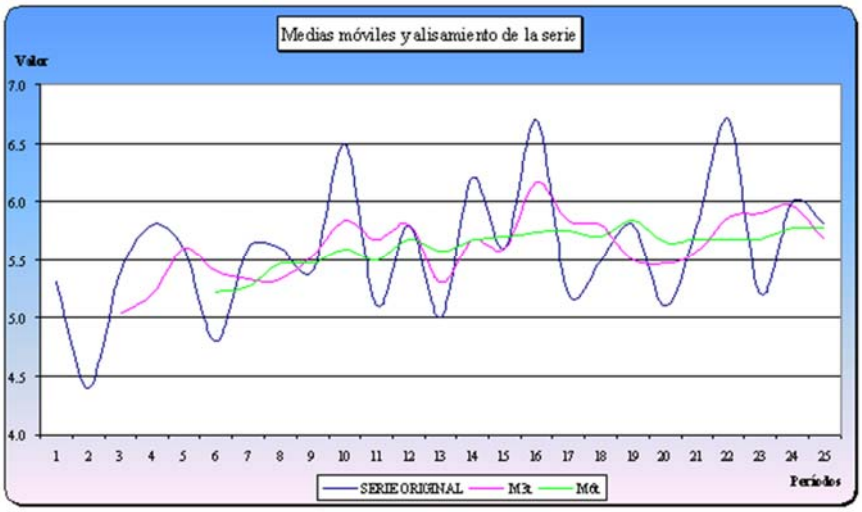


Figura 2.5. Medias móviles y alisamiento exponencial

La expresión general de una media móvil de orden S consistiría en calcular una serie que para cada momento t toma el siguiente valor:

$$M_t = \frac{y_t + y_{t-1} + \Lambda + y_{t-s+1}}{s}$$

La técnica se aplica óptimamente en series sin tendencia ni estacionalidad.

Pronóstico para serie sin tendencia ni estacionalidad por medias móviles

En ese caso, la predicción se realiza con la última media móvil calculada, es decir:

$$\hat{y}_{t+1} = M_t$$

Cuando hay tendencia, debe corregirse el efecto del incremento medio anual; es decir, para predecir valores mensuales:

$$\hat{y}_{t+1} = M_t + \frac{y_t - y_{t-12}}{12}$$

Este resultado es similar a calcular

$$\hat{y}_{t+1} = y_t + \Delta\hat{y}_{t+1} = y_t + M_t = y_t + \frac{y_t - y_{t-12}}{12}$$

Cuando la serie presenta estacionalidad, la media móvil debe ser corregida por coeficientes de estacionalidad. Para esto hay versiones alternativas que consisten en atribuir ponderaciones distintas a cada valor incluido en la media móvil (normalmente más ponderación con el dato más próximo); es decir, calcular *medias móviles ponderadas* y *dobles medias móviles.* No existen criterios generales sobre el orden adecuado de la media móvil.

Alisado exponencial sin tendencia: el alisado simple

Este método consiste en una media móvil con ponderaciones decrecientes en forma de progresión geométrica:

$$M_t = \alpha \cdot y_t + \alpha \cdot (1-\alpha)^1 \cdot y_{t-1} + \alpha \cdot (1-\alpha)^2 \cdot y_{t-2} + \cdots$$

donde α es una constante que varía entre cero y uno, y la suma de los coeficientes de ponderación es uno.

El valor otorgado a la constante α determina el número de términos de la media móvil. Una vez elegido el valor de α y calculada la serie, se puede pasar a realizar predicción. Tal y como se ha descrito, esta técnica sólo es utilizable cuando la serie no presenta tendencia ni estacionalidad.

¿Cuál deber ser la constante α?

Al coeficiente α se le debe atribuir un determinado valor comprendido entre cero y uno, y su elección tiene consecuencias. En primer lugar, si el coeficiente está entre cero y la unidad, entonces los coeficientes que acompañan a cada valor de la serie original en el cálculo de la serie transformada, se convierten en coeficientes de ponderación cuya suma vale uno. Se demuestra que:

$$\alpha + \alpha \cdot (1-\alpha)^1 + \alpha \cdot (1-\alpha)^2 + \cdots = \alpha \sum_{s=0}^{\infty} (1-\alpha)^s = \frac{\alpha}{1-(1-\alpha)} = 1$$

Ello implica que cuanto mayor sea el valor del coeficiente α, menor será el número de términos incluido en la media móvil. Cuando alfa=1, el valor de la media móvil coincide con el valor de la serie en el período. Cuando α se aproxima a cero, las ponderaciones son muy pequeñas y, por tanto, se incluyen gran número de términos. La relación entre el número de términos y el valor del coeficiente α es aproximadamente: $s = \frac{(2-\alpha)}{\alpha}$

La elección del parámetro α depende de las características de la serie objeto de estudio. Hoy en día, los programas de ordenador permiten el cálculo automático del valor óptimo de α, en el sentido de elegir aquel que minimiza el error cuadrático medio. En general, se considera que un α alto es indicativo de fuertes oscilaciones o de tendencia en la serie, lo que conlleva un reducido alisamiento para un mejor ajuste a esas oscilaciones. Por el contrario, para una serie con pequeñas oscilaciones irregulares se aconseja un α reducido (entre 0,01 y 0,4) que implica un fuerte alisado de la serie, al incluir un elevado número de términos.

Pronóstico con alisado exponencial

Como fórmula de predicción la media deberá empezar a calcularse comenzando por el último dato disponible, es decir:

$$\hat{y}_{t+1} = \alpha \cdot y_t + \alpha \cdot (1-\alpha)^1 \cdot y_{t-1} + \alpha \cdot (1-\alpha)^2 \cdot y_{t-2} + \cdots$$

por sustituciones sucesivas, se puede llegar a una expresión alternativa resumida:

$$\hat{y}_{t+1} = \alpha \cdot y_t + (1-\alpha) \cdot \hat{y}_t$$

lo que permite interpretar la predicción con alisado como una media ponderada de los valores previos anteriores reales y de predicción.

También se puede expresar la predicción en función del término de error $e_t = y_t - \hat{y}_t$:

$$\hat{y}_{t+1} = \alpha \cdot y_t + (1-\alpha) \cdot \hat{y}_t = \alpha \cdot (\hat{y}_t + e_t) + (1-\alpha) \cdot \hat{y}_t = \alpha \cdot e_t$$

lo cual nos muestra que las variaciones del valor de predicción $\hat{y}_{t+1} - \hat{y}_t$ son una proporción, α, del error del período anterior:

$$\hat{y}_{t+1} - \hat{y}_t = \alpha \cdot e_t$$

Ello implica que las predicciones sucesivas serán necesariamente iguales, al no disponer de los correspondientes errores y suponerse un valor nulo.

De todas las expresiones anteriores a efectos de cálculo suele utilizarse

$$\hat{y}_{t+1} = \alpha \cdot y_t + (1-\alpha) \cdot \hat{y}_t$$

que exige disponer de un valor para α y de un valor inicial de $\hat{y}_t$. Para esto último se suele adoptar el primer valor de la serie, $\hat{y}_1 = y_1$, o utilizando la media de un número reducido de las primeras observaciones.

¿Cómo influye la tendencia en las medias móviles?

El alisado exponencial, como cualquier otra media móvil sea ponderada o no, sólo es utilizable cuando la serie económica original no presenta tendencia, lo cual no suele ocurrir en la mayor parte de las series económicas.

En el caso de una media móvil simple (sin ponderación), los valores estimados estarán, por ejemplo, sistemáticamente por debajo de los reales cuando existe una tendencia creciente. Para la misma serie, pero utilizando alisado exponencial, el valor óptimo de α será la unidad, ya que, aunque incorporando un sesgo sistemático, éste sería el más reducido posible. Calculando los valores de predicción históricos con $\alpha = 1$, resulta que $\hat{y}_t = y_{t-1}$, con el correspondiente sesgo en una serie con tendencia. Así, para una serie tal como los diez primeros enteros, una media móvil de orden 3 y el mencionado alisado darían los resultados de la Figura 2.6.

y_t	M_t	M_t^3
1	3	---
2	1	---
3	2	2
4	3	3
5	4	4
6	5	5
7	6	6
8	7	7
9	8	8
10	9	9
---	10	9
---	10	9,3
---	10	9,4
---	10	9,3
---	10	9,3

Figura 2.6

En la Figura 2.7 puede verse la evolución real y de predicción (que se corresponde con los datos en cursiva de la tabla), con la característica estabilidad de las predicciones al nivel del último dato de la serie y con la posible perturbación inicial del valor seleccionado para el primer dato de alisado (en este caso $\hat{y} = 3$, media de los cinco primeros datos de la serie original). El problema es similar ante un cambio en el nivel de la serie. Como se aprecia en la Figura 2.8, con una media móvil simple, se centra el sesgo en los periodos inmediatos al cambio.

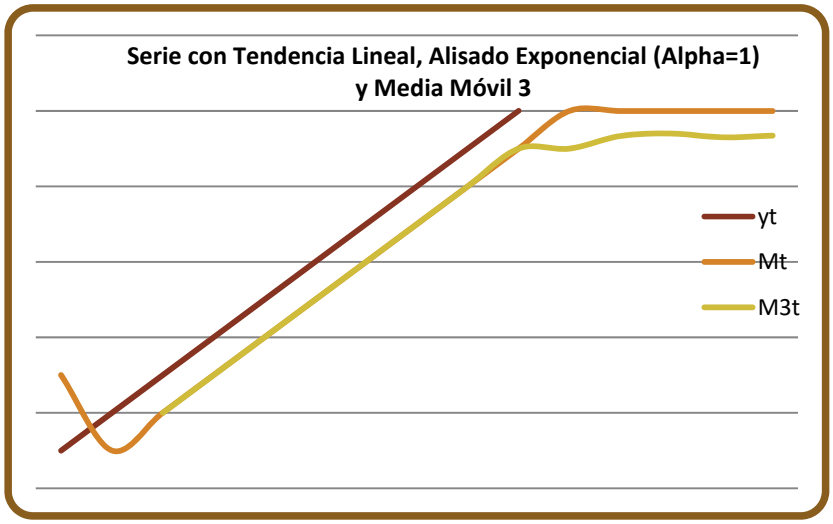


Figura 2.7

y_t	M_t	M^3_t
3	3	
3	3	
3	3	3
3	3	3
3	3	3
3	3	3
6	3	4
6	6	5
6	6	6
6	6	6
	6	*6.0*
	6	*6.0*
	6	*6.0*
	6	*6.0*
	6	*6.0*

Figura 2.8

Con un alisado exponencial (nuevamente, el valor óptimo será $\alpha = 1$, para adaptarse lo más rápidamente posible a la variación) el gráfico de la Figura 2.9 muestra sesgo sólo en el periodo de cambio.

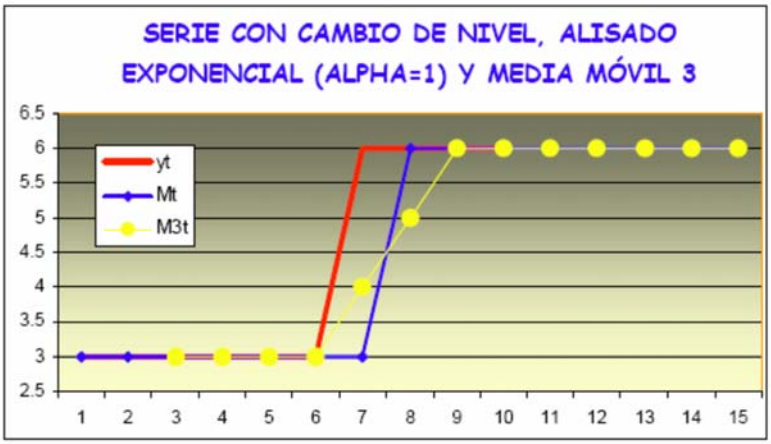


Figura 2.9

Para una serie con estacionalidad, el resultado es relativamente sorprendente. Valores de α cercanos a la unidad, proporcionan una serie histórica estimada, con un comportamiento similar a la real, aunque desfasada un período. Sin embargo, el resultado óptimo, al presentar un ECM mínimo, puede dar un $\alpha = 0$, lo que supone un valor constante de predicción y unas diferencias predicción/realización menores que las que suponen correcciones sistemáticamente atrasadas. Esta situación es calificada por algunos especialistas de *sobrealisado*, al llegarse a un alisado máximo por una inadecuación del propio método para recoger estas oscilaciones sistemáticas.

En las Figuras 2.10 y 2.11 se ha representado una serie supuestamente trimestral con valores iguales cada cuatro periodos y sus predicciones de alisado para $\alpha = 0{,}001$ y $\alpha = 1$.

En resumen, la variante hasta aquí expuesta de alisado exponencial deberá aplicarse a una serie sin tendencia (o eliminada previamente ésta, por ejemplo, tomando diferencias de los valores iniciales); no recoge variaciones estacionales (aunque, posteriormente, puede incorporarse un esquema adicional para su tratamiento); además, sólo proporciona predicciones estables a más de un periodo, por lo que interesará revisar periódicamente las predicciones cada vez que se disponga de un nuevo dato.

y_t	M_t^1	$M_t^{0.001}$
2	2.4	2.4
4	2	2.4
1	4	2.4
3	1	2.4
2	3	2.4
4	2	2.4
1	4	2.4
3	1	2.4
2	3	2.4
4	2	2.4
1	4	2.4
3	1	2.4
	3	*2.4*
	3	*2.4*
	3	*2.4*

Figura 2.10

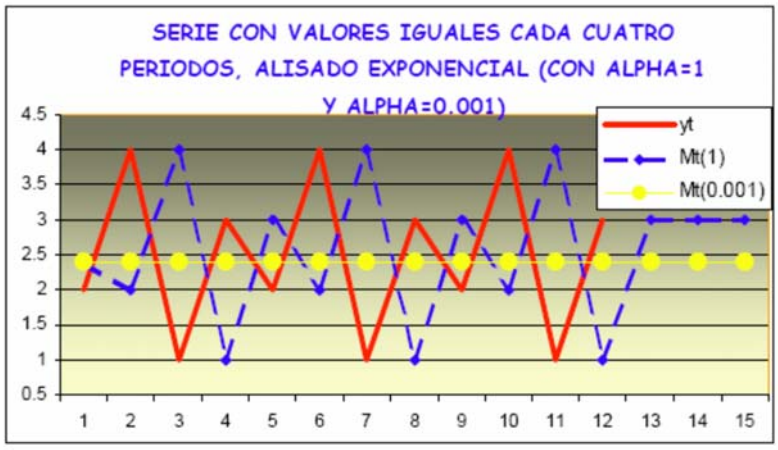


Figura 2.11

Alisados con tendencia

Para las series que presentan tendencia, el desfase sistemático entre valores históricos y media móvil puede corregirse. Pero a efectos de predicción óptima, se necesitan alternativas como el alisado exponencial doble de Brown (también conocido como alisado doble de Brown o alisado exponencial lineal con parámetro único) o el alisado de Holt-Winters (conocido también como alisado exponencial lineal con

doble parámetro), que eliminan el sesgo de la predicción de una serie con tendencia.

Se puede demostrar que en una serie con tendencia los errores históricos (diferencia entre valores de la serie y su media móvil) se igualan a las diferencias entre la media móvil y la doble media móvil. Es decir, si se define la media móvil de orden s como:

$$M_t^s = \frac{\sum_{t-i}^{t-s+1} y_t}{s}$$

y la doble media móvil como:

$$M_t^{s.s} = \frac{\sum_{t-i}^{t-s+1} M_t^s}{s}$$

se tiene que se igualan los errores de ambos, es decir,

$$e_t = y_t - M_t^s = M_t^s - M_t^{s \cdot s} = e_t'$$

Por ello, para obtener una estimación sin sesgo sistemático de la variable y_t se utiliza:

$$M_t^s + e_t' = M_t^s + (M_t^s - M_t^{s \cdot s}) = 2M^s{}_t - M_t^{s \cdot s}$$

Ahora bien, esto tan sólo corrige el desfase sistemático existente en los valores históricos de la serie, y no el funcionamiento de cara a la predicción, ya que los valores predichos se generan sin tendencia. Por ello, se utilizan como alternativas la propuesta por Brown (Alisado exponencial doble) o por Holt-Winters.

El alisado exponencial doble de Brown

Esta técnica (también conocida como alisado exponencial con parámetro único) es adecuada cuando la serie presenta una tendencia

lineal. En ese caso se puede demostrar que se obtiene predicciones óptimas con

$$\hat{y}_{t+h} = a_t + b_t \cdot h$$

donde

$$a_t = M_t^s + (M_t^s - M_t^{s \cdot s}); \quad b_t = \frac{2}{s-1}(M_t^s - M_t^{s \cdot s})$$

Este planteamiento general fue adoptado por Brown en el caso del alisado exponencial. Definido el doble alisado exponencial como:

$$M_t^{\alpha} = \alpha y_t + (1-\alpha) M_{t-1}^{\alpha}$$
$$M_t^{\alpha \cdot \alpha} = \alpha M^{\alpha}{}_t + (1-\alpha) M_{t-1}^{\alpha \cdot \alpha}$$

la ecuación de predicción vendría dada por:

$$\hat{y}_{t+h} = a_t + b_t \cdot h$$

donde

$$a_t = M_t^{\alpha} + (M_t^{\alpha} - M_t^{\alpha \cdot \alpha}); \quad b_t = \frac{\alpha}{1-\alpha}(M_t^{\alpha} - M_t^{\alpha \cdot \alpha})$$

expresión que elimina el sesgo de la predicción de una serie con tendencia.

Es habitual tomar como valor inicial de la media el primer valor de la variable

$$M_1^{\alpha} = M_1^{\alpha \cdot \alpha} = y_1$$

o bien una media de los primeros valores, por ejemplo

$$M_1^{\alpha} = M_1^{\alpha \cdot \alpha} = 1/3(y_1 + y_2 + y_3)$$

Un parámetro α elevado es habitual en series de acusada tendencia. Lo más aconsejable es optar por el α elegido por el programa, normalmente aquel que minimiza el error cuadrático medio.

La técnica Holt-Winters

Se trata de una variante, también conocida como alisado exponencial lineal con doble parámetro, que consigue eliminar el sesgo de la predicción de una serie con tendencia, a través de la inclusión en la media móvil de un componente de tendencia:

$$M_t = \alpha y_t + (1-\alpha)(M_{t-1} + b_{t-1})$$

donde *b* es un factor de variación definido a partir de otra nueva constante de alisado para la tendencia, β :

$$b_t = \beta(M_t - M_{t-1} + (1-\beta)b_{t-1})$$

de forma que la ecuación de predicción adoptaría la forma:

$$\hat{y}_{t+h} = M_t + b_t \cdot h$$

Existen varias posibilidades para la adopción de los valores iniciales. La más sencilla consistiría en hacer

$$M_1 = y_1; \quad b_1 = 0$$

Alternativamente, se puede tomar

$$b_1 = [(y_2 - y_1)/2] + [(y_4 - y_3)/2]$$

o también

$$M_2 = y_2; \quad b_2 = y_2 - y_1$$

Igualmente es aconsejable optar por los valores de los parámetros seleccionados por el programa, en general, elegidos como los que

minimizan los errores cuadráticos medios. Como pauta general, los parámetros α y β tomarán valores elevados (por encima de 0,7) en series de acusada tendencia.

Obviamente, en el caso particular en el que $\beta = 0; \quad b_1 = 0$ el alisado con doble parámetro queda reducido al alisado exponencial simple.

Otros ajuste con funciones matemáticas

Este procedimiento permite tanto calcular el componente tendencia de una serie de forma alternativa, como su aprovechamiento a efectos de predicción. El procedimiento es útil en series con tendencia, pero sin estacionalidad. Consiste en ajustar a los datos una función matemática del tiempo.

Son varias las formulaciones matemáticas posibles: recta, parábola, exponencial, entre otras. Estos ajustes son similares a los ya presentados, sólo que ahora las variables explicativas pueden ser funciones no lineales del tiempo.

En general el cálculo de estos ajustes es sencillo. La dificultad estriba en la elección de la fórmula funcional más adecuada para describir el fenómeno. Una vez obtenidos los coeficientes de la formulación elegida, la predicción se obtiene por la simple sustitución de t por t +1, t + 2,..., t + h.

Es oportuno recordar que, inicialmente, se consideró a las técnicas de las medias móviles o de alisados como aptas para obtener el componente tendencia de una serie económica. La principal característica distintiva de algunos de los más utilizados ajustes de tendencia es que en aquellos casos se obtiene una tendencia variable en el tiempo, eso sí, tanto más estable cuanto mayor sea el número de términos incluidos en la media móvil, o cuanto mayor sea la memoria del proceso de alisado. Como alternativa a estos métodos se pueden utilizar los ajustes de tendencia mediante funciones matemáticas, más o menos complejas, del tiempo.

Fórmulas alternativas de ajuste de tendencia, transformación lineal e instrucciones para su estimación en EViews

Fórmula de ajuste	Transformación lineal	Instrucciones EViews
Recta $y_t = a + bt$	--	LS Y C T
Parábola de 2º grado $y_t = a + bt + ct^2$	$y_t = a + bt + ct_2$ donde $t_2 = t^2$	GENR T2=T^2 LS Y C T T2
Exponencial $y_t = ab^t$	$\text{Log } y_t = \log a + t \log b = c + dt$	GENR LY=LOG (Y) LS LY C T
Potencial $y_t = at^b$	$\text{Log } y_t = \log a + b \log t = c + b \log t$	GENR LY=LOG(Y) GENR LT=LOG(T) LS LY C LT
Exponencial modificada simple $y_t = a + br^t$; $(r<1)$	--	NLS Y=c(1) + c(2)*c(3)^T
Logística $y_t = 1/(a + br^t)$; $(r<1)$	--	NLS Y=1/(c(1)+c(2)*c(3)^T)
Gompertz $\text{Log } y_t = a + br^t$; $(r<1)$	--	GENR LY=LOG(Y) NLS LY=c(1)+c(2)*c(3)^T

Figura 2.12

2.2 Predicción en series con componente estacional

La estacionalidad en una serie presenta características que exigen un tratamiento singular de cara a la predicción. Los procedimientos de predicción de una serie con estacionalidad pasan por aislar tal componente.

Para la obtención del componente de estacionalidad se procede a *separar el componente tendencial*, y a continuación se aísla el componente estacional. Por ello, las técnicas de predicción que se verán son una continuación del proceso previo de desestacionalización.

Entre los procedimientos disponibles para la obtención de la estacionalidad (y de la posterior predicción) se disponen: *la técnica de relación a la media móvil (o Census X-11) y la de Holt Winters.* Para la predicción, se procede aplicando los coeficientes estacionales,

calculados de una u otra forma, sobre la tendencia determinada (en el caso de X-11), o multiplicándolos por la fórmula inicial de Holt-Winters.

Obtención de la tendencia. El primer paso para la descomposición de la serie en tendencia y estacionalidad consiste en calcular la tendencia de la serie original, separando el movimiento regular a largo plazo del conjunto de oscilaciones de la serie. Para ello se dispone de varias alternativas

a) Calculando las diferencias entre valores de la serie original (diferenciación sucesiva);
b) Ajuste de funciones matemáticas, y
c) Media móvil simple, ponderada, o alisado exponencial.

OBSERVACIÓN. De entre estas alternativas resta por analizar con detalle la primera. Consiste en obtener una nueva serie, por ejemplo zt, (que ya no presente componente estacional) a partir de la serie original yt, en la que para cada momento del tiempo t la serie zt toma el siguiente valor:

$$z_t = y_t - y_{t-1} = \Delta y_t$$

Si al representar gráficamente esta serie, zt, se comprueba que oscila alrededor del mismo valor, entonces se trata de la serie original sin tendencia. Si zt crece o decrece a largo plazo, entonces aún presenta tendencia y habría que volver a diferenciar, es decir:

$$\Delta z_t = z_t - z_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = \Delta^2 y_t$$

y, en caso necesario, se seguirá diferenciando, teniendo en cuenta que en cada diferenciación se pierde una observación. En cualquier caso, lo habitual es que con una primera diferencia la serie ya fluctúe en torno al mismo valor, es decir, ya no presente tendencia y, como mucho, se realizaría una segunda diferencia.

Una vez eliminada la tendencia de la serie original, se trata ahora de distinguir, aislando, el patrón de oscilación estacional más común. Para ello, se destacan dos métodos: Relación a la media móvil o CensusX-11 y Holt-Winters.

Census X-11

La técnica de *relación a la media móvil*, y la versión más refinada conocida como Census X-11, es uno de los procedimientos de más difusión ya que, a su propio planteamiento se une el desarrollo de un software adecuado de gran extensión, y que todavía goza de una amplia utilización.

El esquema original de la técnica puede resumirse en tres fases:

1) Obtención del componente tendencia mediante una media móvil. Por ejemplo, para datos mensuales se sugiere una media móvil de 12 términos, como valor de esa tendencia para la media de ese período móvil:

$$M_t^{12} = \frac{1}{12}(y_{t-6} + y_{t-5} + \cdots + y_t + y_{t+1} + \cdots + y_{t+5})$$

Puesto que se trata de un número par de términos, se promedia con la media siguiente:

$$\overline{M}_t^{12} = \frac{1}{2}\left(M_t^{12} + M_{t+1}^{12}\right) = \frac{1}{12}\left(\frac{1}{2}y_{t-6} + y_{t-5} + \cdots + y_t + y_{t+1} + \cdots + y_{t+5} + \frac{1}{2}y_{t+6}\right)$$

2) Establecimiento de los índices o porcentajes de los datos originales respecto a las medias móviles centradas correspondientes:

$$(y_T / \overline{M}_t^{12}) * 100$$

3) Cálculo de la media para cada mes, de estos porcentajes (de todos los eneros, de todos los febreros…) que se tomarán como coeficiente de estacionalidad del esquema multiplicativo:

$$s_i = \frac{1}{n}[(y_{1i} / \overline{M}_{1i}^{12}) * 100) + (y_{2i} / \overline{M}_{2i}^{12}) * 100) + \cdots + (y_{ni} / \overline{M}_{ni}^{12}) * 100)], \quad i = 1,\ldots,12$$

donde el primer subíndice hace referencia al año (n en total) y el segundo al mes de referencia, bien entendido que debe cumplirse que:

$$\sum_{i=1}^{12} s_i = 1200$$

Entre los perfeccionamientos de la técnica propuestos por Shiskin en el *Bureau of the Census* de los EE.UU. en la década de 1950, está el de considerar como una primera estimación estos coeficientes de estacionalidad anteriormente calculados**.** *En este caso se añaden aún tres nuevas etapas.*

4) Cálculo de los valores desestacionalizados de la serie original:

$$y_{ti}^{s} = \frac{y_{ti}}{s_i} * 100$$

5) Establecimiento de una nueva media móvil sobre estos datos desestacionalizados. Se suele proponer una media simple de orden 5 o ponderada de Spencer de orden 15.

6) Con estos nuevos datos se repite el cálculo de índices para cada período y, finalmente, se determinan los coeficientes estacionales definitivos para cada mes.

En el *software* informático X-11 se añaden otras posibilidades de corrección de los datos, como por número de días laborables, por festividades, huelgas, etc.

En cualquier caso, se obtiene coeficientes estacionales que son constantes en el tiempo para cada mes. El mismo procedimiento sirve para obtener, junto a los coeficientes de estacionalidad, la propia serie desestacionalizada.

De cara a la predicción se usará la serie desestacionalizada (pero con tendencia) a la que se le aplicarán cualquiera de los métodos adecuados (alisado exponencial doble, Holt-Winters con dos parámetros,...). La predicción así obtenida debe ser afectada por los indicies de estacionalidad obtenidos por relación a la media móvil o X-11, obteniendo la predicción final.

Holt-Winters

La técnica de Holt-Winters en su variante con tratamiento de la estacionalidad, también conocida como alisado exponencial con triple parámetro, calcula igualmente un índice de estacionalidad para cada dato por relación a la media móvil (y_t / M_t), pero no se intenta establecer un coeficiente medio significativo, sino un procedimiento de actualización de esa estacionalidad estimada. Para ello, y con datos mensuales, la formulación utilizada es:

$$s_t = \delta(y_t / M_t) + (1-\delta) \cdot s_{t-12}$$

expresión que se añade a las dos anteriormente utilizadas en la variante de Holt-Winters para series sólo con tendencia, es decir:

$$M_t = \alpha(y_t / s_{t-12}) + (1-\alpha) \cdot (M_{t-1} + b_{t-1})$$

$$b_t = \beta(M_t - M_{t-1}) + (1-\beta) \cdot b_{t-1}$$

con el único cambio de sustituir en el primer sumando del alisado y_t por y_t / s_{t-12}

A efectos de predicción, se multiplica la fórmula inicial de Holt-Winters por el factor de estacionalidad.

$$\hat{y}_{t+h} = (M_t + b_t h) \cdot s_{t-12+h}$$

2.3 ¿Qué técnica utilizar?

Las técnicas de medias móviles y el alisado exponencial simple, así como los métodos *naïve*, sólo son utilizables en series sin tendencia y sin estacionalidad. Para aplicar estas técnicas a otras series es necesario haber eliminado previamente ambos componentes. Su aplicación se centra en el corto plazo.

Posibilidades de aplicación de técnicas elementales de predicción en función de los componentes de la serie

Serie	Aplicación directa			Alternativa		
Con tendencia y estacionalidad	AHW3P			*Eliminación de tendencia, desestacionalización y posterior afectación del índice de estacionalidad:* ING	MM	AES
				Desestacionalización y posterior afectación del índice de estacionalidad: AED	AHW2P	AFM
Con estacionalidad	AHW3P			*Desestacionalización y posterior afectación del índice de estacionalidad:* ING	MM	AES
Con tendencia	AED	AHW2P	AFM	*Eliminación de tendencia* ING	MM	AES
Sin tendencia ni estacionalidad	ING	MM	AES			

ING=Métodos ingenuos o *naï ve*; MM=Medias móviles; AES=Alisado exponencial simple; AED=Alisado exponencial doble (o de Brown); AHW2P=Alisado Holt-Winters con doble parámetro; AFM=Ajuste de funciones matemáticas; AHW3P=Alisado Holt-Winters con triple parámetro.

Figura 2.13

El alisado exponencial doble (o de Brown), el alisado Holt-Winters con dos parámetros y el ajuste de funciones matemáticas, puede emplearse en series con tendencia pero sin estacionalidad. Si se parte de una serie desestacionalizada, es habitual afectar las predicciones por los factores de estacionalidad. Tan sólo el ajuste de funciones matemáticas podría utilizarse para predicción a medio plazo o incluso, en algunas expresiones, para el largo.

El alisado exponencial Holt-Winters con tres parámetros puede aplicarse directamente a series con tendencia y estacionalidad. Es recomendable su utilización sólo para predicción a corto plazo.

3. ANALISIS BIVARIADO PARA VARIABLES CUANTITATIVAS

El grado de asociación entre dos variables cuantitativas se establece mediante el coeficiente de correlación. Cuando se desea estudiar el efecto de la variación de una (o varias) variable(s) sobre la variación de otra(s), se aplica el análisis de regresión. En este cuaderno se desarrolla el análisis de regresión simple dejando para el próximo cuaderno el desarrollo en profundidad de los métodos econométricos multivariados. Además, cuando las unidades de observación son temporales se puede modelar el comportamiento pretérito de la variable; para ello se desarrolla una introducción al estudio de modelos de series de tiempo.

3.1 Análisis de Correlación

La correlación mide el grado de asociación entre variables cuantitativas. La correlación muestral refleja la tendencia para que los puntos se agrupen sistemáticamente alrededor de una línea recta que crece o decrece de izquierda a derecha; se observa gráficamente al representar dos variables sobre una gráfica bidimensional denominado Diagrama de Dispersión.

La Figura 3.1 muestra la dispersión de las variables PBI a nivel nacional e Índice de Evolución Económica de Río Cuarto (INEVE), las mismas presentan una asociación estadística positiva. Una correlación positiva, refleja una tendencia de alto valor para una primera variable que está asociada con un alto valor de una segunda. Una correlación negativa, refleja una asociación entre un valor alto de una primera variable y un valor bajo de una segunda variable.

El coeficiente de correlación muestral se define como,

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n-1)s_x s_y}; \qquad -1 \leq r \leq 1$$

El valor del coeficiente de correlación cercano a +1, indica una asociación positiva perfecta ente las dos variables; mientras que, si el coeficiente de correlación es cercano a -1, existe una asociación negativa perfecta. Un coeficiente de correlación cercano a 0, refleja la ausencia de asociación lineal.

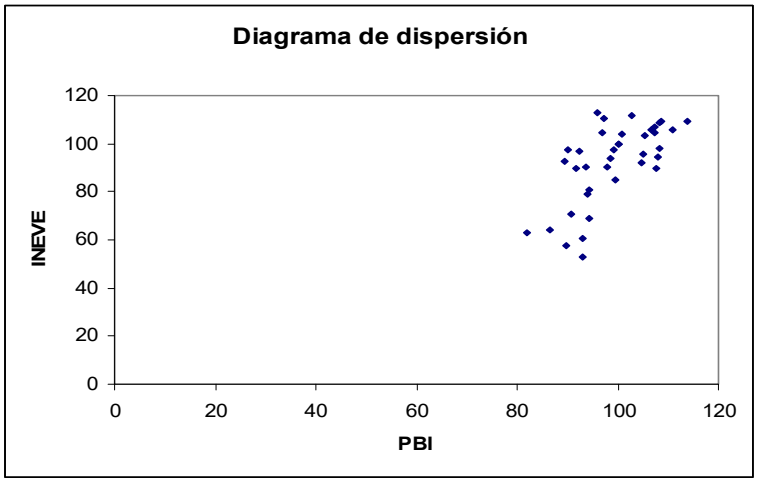


Figura 3.1 Diagrama de dispersión

Ejemplo 3.1. La varianza observada en el indicador local de actividad económica (INEVE) es de 274,55 y la varianza en el indice de PBI 56,01. La covarianza entre ambas variables es de 71,56. ¿Cuál es la correlación entre ambas variables?

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n-1)s_x s_y} = \frac{Cov(X,Y)}{\sqrt{S_X^2}\sqrt{S_Y^2}} = \frac{71{,}56}{\sqrt{274{,}55}\sqrt{56{,}01}} = 0{,}5927$$

3.2 Introducción al análisis de regresión

En el amplio espectro del estudio de las relaciones entre variables, tanto cuantitativas como cualitativas, que involucra el proceso de investigación econométrica, existe la posibilidad de establecer relaciones funcionales. En algunas de ellas, la variación de una variable dependiente es explicada por las variaciones de una o más variables independientes, mediante funciones lineales que involucran adicionalmente un término aleatorio. En particular, en este capítulo se inicia el estudio de las relaciones lineales entre dos variables cuantitativas. Para ello, se estudia la forma de especificación y estimación de esas relaciones lineales y la bondad de las mismas.

Especificación del modelo

En el análisis de la información existen, frecuentemente, un número de variables claves que se convierten en el centro de atención del estudio. El análisis de regresión proporciona una herramienta que puede cuantificar tales relaciones y proporcionar un control estadístico. Es un análisis de dependencia puesto que involucra una variable dependiente, como el punto de atención del análisis, la cual es explicada por las variables independientes o variables explicativas.

El análisis de regresión está orientado hacia

- la descripción, obtención de las relaciones entre variables independientes y una variable dependiente para un conjunto de observaciones.
- la predicción, niveles a alcanzar por la variable dependiente bajo el supuesto de comportamiento de la variable explicativa.

La construcción de un modelo de regresión, generalmente, empieza con la especificación de la variable dependiente y de la variable, o variables, independientes.

$$Y = f(X)$$

Y : variable dependiente o endógena

X : variable independiente o exógena.

Como no se puede explicar en forma perfecta el comportamiento de Y sólo por X, se adiciona un término de error, ε:

$$Y = f(X) + \varepsilon$$

A ésta se la denomina *ecuación de regresión* de Y sobre X. ε es el residuo o error y es una variable aleatoria. Este término hace referencia a error de medición de las variables, error de especificación del modelo y omisión de otras variables explicativas que influyen sobre el comportamiento de Y. Al ser ε variable aleatoria, también lo es Y.

Se explicita una relación lineal para $f(X)$

$$f(X) = \alpha + \beta X$$

Si se dispone de n observaciones de Y y de X, puede especificarse el siguiente modelo de regresión.

$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad i = 1, \Lambda, n$$

donde,

Y = variable dependiente

X = variable explicativa

ε= perturbación aleatoria o término de error

α, β = parámetros del modelo.

El modelo está basado en las siguientes consideraciones:

La relación especificada es lineal. Esto conduce a tener que linealizar aquellas relaciones funcionales no lineales a efectos de poder aplicar el análisis.

El término de error es una variable aleatoria donde:

- El promedio es cero.

$$E(\varepsilon_i) = 0 \quad \forall i$$

- La varianza es constante. Esto significa que el error no es más grande para valores grandes de X que para valores pequeños de X. Esto se lo conoce como “homocedasticidad”.

$$V(\varepsilon_i) = E(\varepsilon_i \varepsilon_j) = \sigma^2 \quad \forall i = j$$

- Los errores son independientes entre sí. Esto significa que los errores de una observación no dependen de la observación anterior. Esto se lo conoce como “no autocorrelación”.

$$Cov(\varepsilon_i \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0 \quad \forall i \neq j$$

- El error tiene distribución normal de media cero y varianza constante.

$$\varepsilon_i \sim N(0, \sigma^2)$$

- La variable explicativa X es fija, no estocástica e independiente de los errores

$$Cov(\varepsilon_i X_i) = E(\varepsilon_i X_i) = 0$$

Bajo el cumplimiento de este conjunto de supuestos, los estimadores minimocuadráticos son lineales, insesgados y óptimos (ELIO) y es posible construir intervalos de confianza y realizar test para $\hat{\alpha}$ y $\hat{\beta}$.

Estimación de los parámetros del modelo

Los parámetros α y β se estiman por el método de "mínimos cuadrados". Esto es elegir $\hat{\alpha}$ y $\hat{\beta}$ como estimadores de α y β, respectivamente, de tal manera que la suma de las diferencias al cuadrado entre el valor observado y el valor estimado de la variable dependiente sea mínima. Es decir, que tenga la siguiente propiedad

$$Q = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = mínimo$$

La gráfica muestra la dispersión de la información proveniente de la observación de dos variables X e Y. La línea recta simula el modelo estimado. La diferencia entre el valor observado y la línea estimada es el error. Lo que se "minimiza" es la suma de cuadrados entre los puntos y la línea, en forma vertical.

Los parámetros α y β son las características de la relación entre X e Y. Reemplazando en Q, $\hat{Y}$ por su igual:

$$Q = \sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right)^2$$

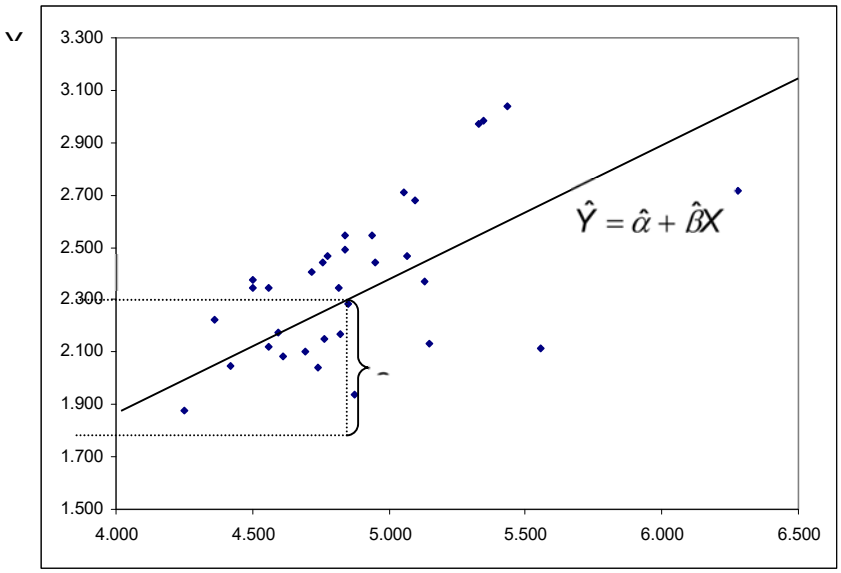


Figura 3.2 Gráfico de dispersión

Para que el valor de Q sea mínimo se debe derivar con respecto a $\hat{\alpha}$ y $\hat{\beta}$ e igualar a cero.

$$\frac{\partial Q}{\partial \alpha} = 2\sum \left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right)(-1) = 0$$

$$\sum Y_i = n\hat{\alpha} + \hat{\beta}\sum X_i$$

$$\frac{\partial Q}{\partial \beta} = 2\sum \left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right)(-X_i) = 0$$

$$\sum X_i Y_i = \hat{\alpha}\sum X_i + \hat{\beta}\sum X_i^2$$

De esta manera tenemos un sistema de dos ecuaciones y dos incógnitas denominadas ecuaciones normales. De la primera se obtiene el valor de $\hat{\alpha}$

$$\sum Y_i = n\hat{\alpha} + \hat{\beta}\sum X_i$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

Reemplazando ésta en la segunda se obtiene $\hat{\beta}$

$$\sum X_i Y_i = \left(\bar{Y} - \hat{\beta}\bar{X}\right)\sum X_i + \hat{\beta}\sum X_i^2$$

$$\sum X_i Y_i = \left(\bar{Y} - \hat{\beta}\bar{X}\right)n\bar{X} + \hat{\beta}\sum X_i^2$$

$$\sum X_i Y_i - n\bar{X}\bar{Y} = \hat{\beta}\sum X_i^2 - \hat{\beta}\bar{X}^2$$

$$\hat{\beta} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2}$$

$\hat{\alpha}$ y $\hat{\beta}$, se denominan coeficientes de regresión y se basan en la muestra aleatoria. Cada muestra aleatoria tendrá asociada un $\hat{\alpha}$ y un $\hat{\beta}$, *diferentes*. Una medida de esta variación de las estimaciones de los parámetros está dada por sus *errores estándar*. De este modo, así como la media muestral, $\bar{X}$, tiene un error estándar que es estimado por $\sigma_{\bar{X}}$; también $\hat{\alpha}$ y $\hat{\beta}$ tienen un error estándar asociado con cada uno de ellos, $s_{\hat{\alpha}}$ y $s_{\hat{\beta}}$.

$$s_{\hat{\alpha}} = \sqrt{\sigma_e^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X-\bar{X})^2}\right]} \qquad s_{\hat{\alpha}} = \sqrt{\frac{\sigma_e^2}{\sum(X-\bar{X})^2}}$$

La varianza de los errores de la regresión, σ_e^2, se estima a partir de s_e^2

$$s_e^2 = \frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{\sum e^2}{n-2}$$

La interpretación de los parámetros estimados

Los parámetros tienen un significado muy preciso.

- El parámetro β indica que, si la variable X cambia en una unidad, la variable Y cambiará en β unidades.
- El parámetro α refleja el nivel que alcanza la variable dependiente cuando la variable explicativa asume el valor nulo.

Tal como se indicó anteriormente, la estimación de $\hat{\beta}$, tiene una variación asociada con ella (medida por $s_{\hat{\beta}}$), porque se basa en una muestra de individuos. Una forma de evaluar la magnitud de $\hat{\beta}$, tomando en cuenta su variación, consiste en usar una prueba estadística de hipótesis. Para esto se calcula el estadístico t,

$$t = \frac{\hat{\beta} - \beta}{s_{\hat{\beta}}}$$

Bajo la hipótesis nula, que será planteada como $\beta = 0$, el estadístico t se reduce a,

$$t = \frac{\hat{\beta} - \beta}{s_{\hat{\beta}}} \sim t_{n-2}$$

Nuevamente, como se hizo en otras pruebas de hipótesis, este valor empírico de t se compara con el valor teórico. Si el valor empírico es superior, en valor absoluto, al valor teórico se rechaza la hipótesis nula y se decide que $\beta \neq 0$ y por lo tanto, la relación entre X e Y es significativa, a un valor de probabilidad α determinado.

Intervalos de confianza para los parámetros

Para construir intervalos de confianza para los parámetros de la regresión, se utiliza la distribución t, teniendo en cuenta que

$$\frac{\hat{\alpha} - \alpha}{s_{\hat{\alpha}}} \sim t_{n-2} \qquad \frac{\hat{\beta} - \beta}{s_{\hat{\beta}}} \sim t_{n-2}$$

Explícitamente, para el parámetro α es

$$P\left(t_{n-2;\alpha/2} \leq \frac{\hat{\alpha} - \alpha}{s_{\hat{\alpha}}} \leq t_{n-2;1-\alpha/2}\right) = 1 - \alpha$$

Resolviendo de igual manera a como se hiciera para estimar por intervalo el valor de μ, se tiene

$$P\left(\hat{\alpha} - s_{\hat{\alpha}} t_{n-2;\frac{\alpha}{2}} \leq \alpha \leq \hat{\alpha} + s_{\hat{\alpha}}\, t_{n-2;1-\alpha/2}\right) = 1 - \alpha$$

Para el parámetro β se procede de igual manera.

También es posible estimar el intervalo de confianza para la varianza de la estimación. Aquí hay que recordar que la varianza se distribuye chi cuadrado y que el intervalo se construye haciendo

$$P\left(\chi^2_{gl;\,\alpha/2} \leq \frac{gl\, s^2}{\sigma^2} \leq \chi^2_{gl;\,1-\alpha/2}\right) = 1 - \alpha$$

De modo que, el intervalo de confianza de la varianza es

$$P\left(\frac{gl\, s^2}{\chi^2_{gl;\,1-\alpha/2}} \leq \sigma^2 \leq \frac{gl\, s^2}{\chi^2_{gl;\,\alpha/2}}\right) = 1 - \alpha$$

Donde gl son los grados de libertad que, en regresión simple, es igual a $n - 2$.

Predicción

El modelo de regresión, desde luego, puede ser usado como una herramienta predictiva. Dada la ecuación de regresión

$$Y = \hat{\alpha} + \hat{\beta} X$$

se pueden predecir valores futuros de Y (Y_F) dado un valor futuro de X (X_F) usando

$$\hat{Y}_F = \hat{\alpha} + \hat{\beta} X_F$$

El valor de Y_F viene dado por

$$Y_F = \alpha + \beta X_F + \mu_F$$

De aquí el error de predicción es

$$\hat{Y}_F - Y_F = (\hat{\alpha} - \alpha) + (\hat{\beta} - \beta)X_F - \mu_F$$

Como

$$\left.\begin{matrix} E(\hat{\alpha} - \alpha) = 0 \\ E(\hat{\beta} - \beta) = 0 \\ E(\mu_F) = 0 \end{matrix}\right\} \rightarrow E(\hat{Y} - Y_F) = 0$$

De donde: $E(\hat{Y}_F) = Y_F$

Esto significa que el predictor $\hat{Y}_F$ es <u>insesgado.</u>

La varianza del error de predicción es:

$$V(\hat{Y}_F - Y_F) = V(\hat{\alpha} - \alpha) + X_F^2 V(\hat{\beta} - \beta) - 2X_F cov(\hat{\alpha} - \alpha, \hat{\beta} - \beta) + V(\mu_F)$$

$$= \sigma_e^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum (X - \overline{X})^2} \right) + X_F^2 \frac{\sigma_e^2}{\sum (X - \overline{X})^2} - 2X_F \sigma_e^2 \frac{\overline{X}}{\sum (X - \overline{X})^2} + \sigma_e^2$$

$$= \sigma_e^2 \left[1 + \frac{1}{n} + \frac{(X_F - \overline{X})^2}{\sum (X - \overline{X})^2} \right]$$

Por lo tanto, $V(\hat{Y}_F - Y_F)$ se incrementa a medida que X_F se aleja de $\overline{X}$. Como σ^2 es desconocida se estima por:

$$s_e^2 = \frac{SCR}{n-2} = \frac{\sum e^2}{n-2}$$

De aquí, el error estándar de la predicción es

$$s_{\hat{Y}_F} = \sqrt{\frac{\sum e^2}{n-2}\left[1+\frac{1}{n}+\frac{\left(X_F-\overline{X}\right)^2}{\sum\left(X-\overline{X}\right)^2}\right]}$$

Esta última expresión se puede utilizar para construir intervalos de confianza para Y_F.

Para tener en cuenta:

- La predicción usando valores extremos de la variable independiente puede ser arriesgada. Recuerde que el supuesto lineal puede ser apropiado solo para un rango limitado de las variables independientes. Además, la muestra aleatoria no proporcionó información acerca de los valores extremos de peso.
- Si el contexto cambia, como el hecho de comunidades diferentes, entonces, los parámetros del modelo se verán afectados. Los datos provenientes de la muestra aleatoria fueron obtenidos bajo un conjunto de condiciones de la población. Si cambian, entonces el modelo puede verse afectado.

Coeficiente de determinación

Para evaluar la capacidad predictiva del modelo se usa el coeficiente de determinación o r^2. Este coeficiente es la razón de la variación explicada sobre la variación total:

$$r^2 = \frac{\text{var}\,iación\quad total\quad -\quad \text{var}\,iación\,no\,\text{exp}\,licada}{\text{var}\,iación\;\;total} = \frac{\text{var}\,iación\quad \text{exp}\,licada}{\text{var}\,iación\,total}$$

$$r^2 = \frac{\sum (Y_i - \overline{Y})^2 - \sum (\hat{Y}_i - Y_i)^2}{\sum (Y_i - \overline{Y})^2} = \frac{\sum (\hat{Y}_i - \overline{Y})^2}{\sum (Y_i - \overline{Y})^2}$$

La variación total en Y se divide en la variación que explica el modelo de regresión, variación explicada, y la variación que no explica el modelo de regresión, variación no explicada.

El término r^2 es el cuadrado de la correlación entre X e Y. Por consiguiente, se encuentra entre cero y uno.

Ejemplo 3.2. En un centro médico quieren estudiar, de la población de hombres adultos aparentemente sanos, la relación existente entre la concentración de glucosa en la sangre con el peso. Específicamente, están interesados en conocer si el peso de las personas determina el nivel de glucosa existente en la sangre.

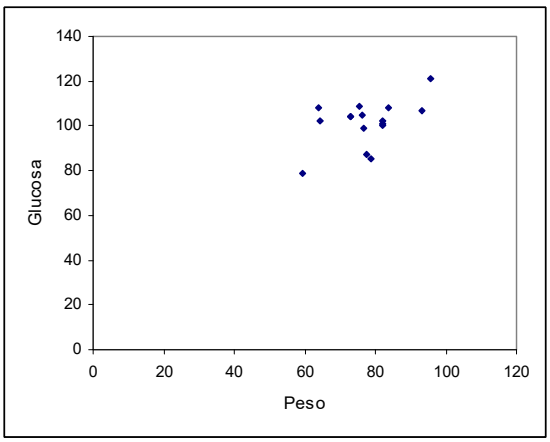


Figura 3.3 Gráfico de dispersión

Observación	Peso (en kgrs.) X	Glucosa (mg/100 ml) Y	XY	X^2	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$
1	64,0	108	6912,0	4096,00	94,520	13,480	181,710
2	75,3	109	8207,7	5670,09	100,283	8,717	75,986
3	73,0	104	7592,0	5329,00	99,110	4,890	23,912
4	82,1	102	8374,2	6740,41	103,751	-1,751	3,066
5	76,2	105	8001,0	5806,44	100,742	4,258	18,131
6	95,7	121	11579,7	9158,49	110,687	10,313	106,358
7	59,4	79	4692,6	3528,36	92,174	-13,174	173,554
8	93,4	107	9993,8	8723,56	109,514	-2,514	6,320
9	82,1	101	8292,1	6740,41	103,751	-2,751	7,568
10	78,9	85	6706,5	6225,21	102,119	-17,119	293,060
11	76,7	99	7593,3	5882,89	100,997	-1,997	3,988
12	82,1	100	8210,0	6740,41	103,751	-3,751	14,070
13	83,9	108	9061,2	7039,21	104,669	3,331	11,096
14	73,0	104	7592,0	5329,00	99,110	4,890	23,912
15	64,4	102	6568,8	4147,36	94,724	7,276	52,940
16	77,6	87	6751,2	6021,76	101,456	-14,456	208,976
Sumas	1237,8	1621	126128,1	97178,60			1204,648

Figura 3.4 Tabla de datos

Las observaciones fueron graficadas en el diagrama de dispersión de la Figura 3.3, donde, cada punto de la gráfica, representa un individuo para el cual se observó el peso y el nivel de glucosa. La dispersión de la nube de puntos presenta una asociación positiva.

El siguiente paso, consiste en obtener una línea que tenga el mejor "ajuste" para estos puntos. La línea se denomina línea de mínimos cuadrados y se obtiene a partir de la especificación del modelo de regresión:

$$Y = \alpha + \beta X + \varepsilon$$

donde:

Y , es la variable dependiente "nivel de glucosa"

X, es la variable explicativa "peso"

α y β , parámetros a estimar

μ , término de error

Se deben calcular los valores de los estimadores de los parámetros α y β a partir de los datos consignados en la tabla. De allí se obtienen los valores promedio de las variables Peso (X) y Concentración de glucosa (Y).

$$\overline{X} = \frac{\sum X}{n} = \frac{1237.8}{16} = 77.3625$$

$$\overline{Y} = \frac{\sum Y}{n} = \frac{1621}{16} = 101.3125$$

Se calculan los coeficientes de la regresión

$$\hat{\beta} = \frac{\sum X_i Y_i - n\overline{X}\overline{Y}}{\sum X_i^2 - n\overline{X}^2} = \frac{126128.1 - 16 * 77.3625 * 101.3125}{97178.6 - 16 * 77.3625^2} = 0.50975$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} = 101.3125 - 0.50975 * 77.3625 = 61.8769$$

Con estos resultados la línea de regresión estimada es

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X = 61.88 + 0.51X$$

El valor 61.88 es una estimación del parámetro α, e indica el nivel de glucosa independiente del peso alcanzado por la persona. El valor 0.51, es una estimación del parámetro β que indica en cuánto cambia el nivel de glucosa cuando cambia el peso en una unidad.

El término $\hat{Y}$ es una estimación del nivel de glucosa basada en el modelo de regresión, por ejemplo, cuando X es 64:

$$\hat{Y} = 61.88 + 0.51 * 64 = 94.520$$

De este modo, para un paciente de 64 Kg., la estimación del nivel de glucosa es de 94.52.

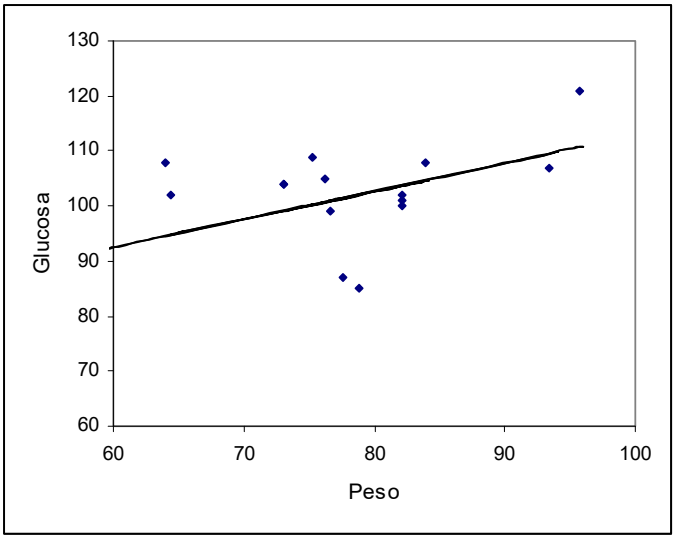


Figura 3.4. Gráfico de dispersión

Si se obtuviera *otra muestra* aleatoria de 16 varones adultos, contendría individuos diferentes. Como resultado, el gráfico de dispersión XY diferiría de la figura anterior y los coeficientes de regresión, $\hat{\alpha}$ y $\hat{\beta}$, serían diferentes.

La varianza de los errores y de los coeficientes se obtiene de la siguiente manera:

$$s_e^2 = \frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{\sum e^2}{n-2} = \frac{1204.648}{14} = 86.0463$$

$$s_{\hat{\alpha}}^2 = \sigma_e^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum\left(X-\bar{X}\right)^2}\right] = \sigma_e^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum X^2 - n\bar{X}^2}\right]$$

$$= 86.0463\left[\frac{1}{16} + \frac{77.3625^2}{97178.60 - 16 * 77.3625^2}\right]$$

$$= 368.2217$$

$$s^2_{\hat{\beta}} = \frac{\sigma^2_e}{\sum\left(X-\overline{X}\right)^2} = \frac{\sigma^2_e}{\sum X^2 - n\overline{X}^2}$$

$$= \frac{86.0463}{97178.60 - 16 * 77.3625^2} = 0.0606$$

Con estos valores los desvíos de la regresión y los coeficientes alcanzan los siguientes valores:

$$s_e = \sqrt{86.0463} = 9.2761$$

$$s_{\hat{\alpha}} = \sqrt{368.2217} = 19.1891$$

$$s_{\hat{\beta}} = \sqrt{0.0606} = 0.2462$$

Hasta aquí se ha realizado una estimación puntual del valor de los coeficientes de la regresión. Pero es posible, como ocurre con otros parámetros, realizar una estimación por intervalo asignándole una probabilidad de ocurrencia.

$$P\left(\hat{\alpha} - t_{n-2} s_{\hat{\alpha}} \le \alpha \le \hat{\alpha} + t_{n-2} s_{\hat{\alpha}}\right) = 1 - 0.05$$
$$P(61.8769 - 2.1448 * 19.1891 \le \alpha \le 61.8769 + 2.1448 * 19.1891) = 0.95$$
$$P(20.7201 \le \alpha \le 103.0337) = 0.95$$

$$P\left(\hat{\beta} - t_{n-2} s_{\hat{\beta}} \le \beta \le \hat{\beta} + t_{n-2} s_{\hat{\beta}}\right) = 1 - 0.05$$
$$P(0.50975 - 2.1448 * 0.2462 \le \beta \le 0.50975 + 2.1448 * 0.2462) = 0.95$$
$$P(-0.01829 \le \beta \le 1.03779) = 0.95$$

Cuando se plantea el modelo de regresión, el primer supuesto es que el comportamiento de la variable X afecta el comportamiento de la variable Y. Para probar esto se contrasta la hipótesis de que el parámetro β es nulo. Si esto fuera así; entonces, no habría efecto del peso sobre los niveles de glucosa y el modelo no debería ser usado para ningún propósito.

$$H_0 : \beta = 0$$

El estadístico de prueba es

$$t = \frac{\hat{\beta} - \beta_{H_0}}{s_{\hat{\beta}}} = \frac{0.50975 - 0}{0.2462} = 2.07$$

Al nivel de significación de 0.05, $\alpha = 0.05$, el valor de la distribución t con 14 grados de libertad (valor teórico de t) es de ± 2.144. Esto significa que el valor de prueba cae en la zona de aceptación de la H_0, por lo que el peso alcanzado por los hombres adultos aparentemente sanos no afecta el nivel de glucosa observado en ellos.

Para predecir el valor del nivel de glucosa a partir del modelo de regresión, se supone un nivel de peso y se reemplaza en el mismo. Si se propone un peso de 80 Kg, un nivel de glucosa estimado en base al modelo, sería:

$$\hat{Y} = 61.88 + 0.51 * 80 = 102.68$$

El r^2 es igual a 0.23; por consiguiente, el 23% de la variación total de Y, está explicada por X

4. ANALISIS BIVARIADO PARA VARIABLES CUALITATIVAS

El grado de asociación entre dos variables cualitativas se estudia mediante el análisis de tablas de contingencia.

4.1. Análisis de Tablas de Contingencia

La Tabla de Contingencia reúne la totalidad de coocurrencias que presenta la población para dos variables consideradas, donde cada variable tiene modalidades mutuamente excluyentes entre sí.

La Tabla de Contingencia que surge de la observación, censal o muestral, se la denomina Tabla de frecuencias observadas. Tanto las filas como las columnas, brindan información sobre las modalidades de las variables. Por esto es importante conocer el total de observaciones en cada modalidad y el peso que éste (el total de la modalidad) tiene en el total de observaciones realizadas en la población. A éste último indicador se lo denomina perfil; así se tiene el perfil columna (peso del total modalidad de la columna en el total de observaciones) y el perfil fila (peso del total modalidad de la fila en el total de observaciones).

Los perfiles (fila y columna) surgidos de una tabla de frecuencias observadas permiten construir la tabla de frecuencias esperada bajo situación de independencia. Esta tabla es teórica y se construye con el producto de los perfiles fila y columna y el total de observaciones. Formalmente

$$p_f \; x \; p_c \; x \; n$$

Donde p_f y p_c son los perfiles fila y columna medidos en número de observaciones y n es la cantidad de observaciones.

El estadístico chi-cuadrado

El estadístico chi-cuadrado es una medida de la diferencia, entre las frecuencias observadas con la frecuencia esperada, bajo el supuesto de independencia estadística. El estadístico de chi-cuadrado, se define como:

$$\chi^2_{(f-1)(c-1)} = \sum \frac{(f_a - f_a^e)^2}{f_a^e}$$

donde, f_a es la frecuencia absoluta observada, f_a^e es la frecuencia absoluta esperada bajo el supuesto de independencia, f es el número de filas, c es el número de columnas y $(f-1)(c-1)$ son los grados de libertad

Si las variables son estadísticamente independientes, el valor χ^2 debería ser relativamente pequeño. Sin embargo, si las variables no son independientes -si están asociadas o relacionadas-, entonces el valor de χ^2 debería ser relativamente grande.

Pruebas de independencia

La hipótesis nula asociada con el estadístico muestral chi-cuadrado, es que las dos variables cualitativas son estadísticamente independientes.

La prueba de la hipótesis se basa en el hecho de que el estadístico chi-cuadrado está distribuido chi-cuadrado con $(f-1)(c-1)$ grados de libertad, suponiendo que la hipótesis nula es verdadera. O, en otras palabras, dado un nivel de significación α se puede determinar, de acuerdo a los grados de libertad, el valor teórico de Chi- Cuadrado, se lo compara con el valor empírico y si el segundo es superior, se rechaza la hipótesis de independencia.

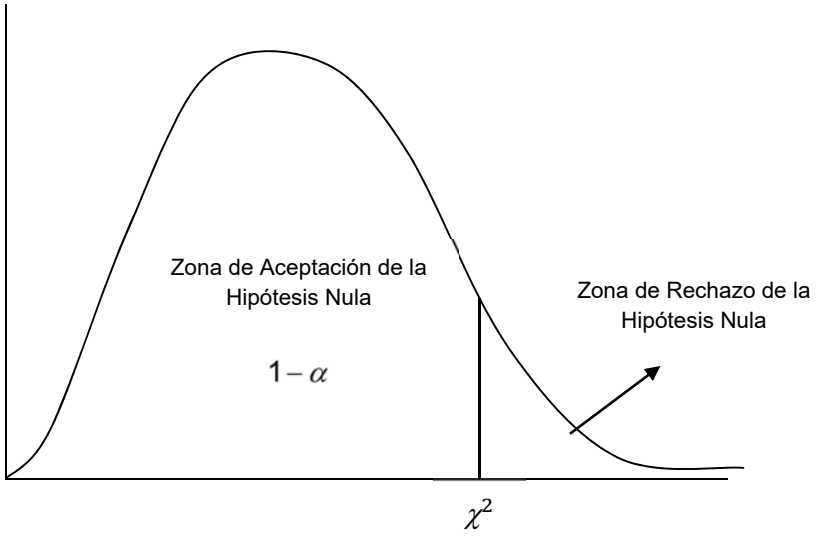


Figura 4.1 Prueba de la hipótesis de independencia

Ejemplo 4.1. La Figura 4.2 muestra los resultados de una encuesta de 200 asistentes a las funciones de ópera que fueron interrogados acerca de la frecuencia con la que asistían a los conciertos de la orquesta sinfónica en una ciudad vecina. La frecuencia de asistencia, fue dividida en las categorías de frecuente, ocasional y nunca; y se les preguntó si consideraban la ubicación de la sinfónica como conveniente o inconveniente.

Asistencia a los conciertos de la sinfónica (A)	Ubicación (U)		Total	Perfil fila p_f
	Conveniente	No conveniente		
Frecuente	22	18	40	0.20
Ocasional	48	52	100	0.50
Nunca	10	50	60	0.30
Total	80	120	200	1.00
Perfil columna p_c	0.40	0.60	1.00	

Figura 4.2 Asistentes a la Opera – Frecuencia Observada

La tabulación cruzada resultante, muestra la clasificación porcentual de asistencia en cada categoría de localización. Los totales de las filas, los totales de las columnas y las proporciones (p_f y p_c), son tabulados al margen. p_f y p_c son las distribuciones de frecuencias marginales para las variables respectivas. Por ejemplo, la fila total indica que 80 entrevistados (0.40 de todos los entrevistados), consideraban que la localización era conveniente y 120 (0.60 de todos los entrevistados) que era inconveniente.

Si el experimento se repitiera, un 20% de los resultados mostrarían que la Asistencia a los conciertos de la sinfónica pertenece a la categoría de "frecuentes". La frecuencia absoluta esperada, bajo la condición de independencia estadística, sería de 0.20n, donde n es el número de nuevos experimentos realizados. También se espera, bajo las mismas condiciones que el 40% de los casos tengan una ubicación "conveniente" y el 60% una ubicación "no conveniente".

De igual manera, el número de experimentos que darían como resultado una asistencia frecuente en una ubicación conveniente “se esperaría” que fueran $0{,}4x0{,}2xn$

Asistencia a los conciertos de la sinfónica (A)	Ubicación (U)		Total	Perfil fila p_f
	Conveniente	No conveniente		
Frecuente	16	24	40	0.20
Ocasional	40	60	100	0.50
Nunca	24	36	60	0.30
Total	80	120	200	1.00
Perfil columna p_c	0.40	0.60	1.00	

Figura 4.3 Asistentes a la Opera – Frecuencia esperada

El valor empírico para el estadístico chi cuadrado es:

$$\chi^2_{(f-1)(c-1)} = \sum \frac{(f_a - f_a^e)^2}{f_a^e} = \\ = \frac{(22-16)^2}{16} + \frac{(18-24)^2}{24} + \frac{(48-40)^2}{40} + \frac{(52-60)^2}{60} \\ + \frac{(10-24)^2}{24} + \frac{(50-36)^2}{36} = 20.02$$

El estadístico de chi-cuadrado teórico, para 2 grados de libertad y nivel de confianza de 0.95, es de 5.991. Este valor crítico divide la zona de aceptación de la hipótesis nula de la zona de rechazo.

El valor empírico de 20.02 es mayor al teórico, esto indica que las dos variables pueden no ser estadísticamente independientes; al menos en este caso particular, se encontró una asociación muestral entre las dos variables.

Droesbesque y Fine (1997) señalan que, para el tratamiento uni o multidimensional de una muestra mediante software estadístico, se supone que todos los individuos tienen el mismo peso y que las observaciones provienen de una muestra independiente e idénticamente distribuida; es decir, la muestra de individuos es una muestra aleatoria simple con reemplazo (o sin reemplazo si el tamaño de la población es muy grande respecto al tamaño de la muestra) y por ende los test estadísticos se fundamentan en esta hipótesis.

Existen softwares estadísticos que permiten el tratamiento de datos de encuestas realizadas con diseño en etapas múltiples y proveen las varianzas de los estimadores. Pero es posible que el investigador construya las tablas, a través de planillas de cálculos, para tratar datos de encuestas sobre muestras complejas.

OBSERVACIÓN. Para el tratamiento de datos que provienen del muestreo estratificado, hay que crear una hoja de trabajo con tantas líneas como estratos y agregar los tamaños de la población y de la muestra, las medias, las varianzas, covarianzas por estrato, y programar las fórmulas. Por ejemplo, para la estimación por intervalo de confianza, de la media de una variable cuantitativa X se tendría que diseñar la hoja de cálculo de la Figura 4.4.

$Estrato\ k$	N_k	n_k	$\bar{x}_k$	s_k^2	$(N_k/N)\ \bar{x}_k$	$(N_k/N)^2 s_k^2/n_k$
1						
:						
K						
$Suma$	N				$\hat{\mu}$	$\hat{V}(\hat{\mu})$

Figura 4.4 Hoja de trabajo para cálculos

Ejemplo 4.2. Droesbesque y Fine (1997) plantean la realización de una encuesta para estimar la proporción de dirigentes satisfechos con una medida gubernamental; paralelamente, se desea conocer las diferencias en la satisfacción según el tamaño de la empresa. La población de referencia son los dirigentes de empresas francesas del sector secundario y terciario, en ellos se observó la actividad (Industria, Comercio, Servicios) y el tamaño de la empresa medido por el número de asalariados (menos de 50, de 50 a 100, de 100 a 500 y más de 500) (Figura 4.5):

	Tamaño				
Sector	-50	50 a 100	100 a 500	+500	Total
Industria	38200	6000	4800	1000	50000
Comercio	43400	4000	2200	400	50000
Servicios	39300	5000	5000	700	50000
Total	120900	15000	12000	2100	150000

Figura 4.5 Clasificación de los dirigentes entrevistados, por sector actividad y tamaño de la empresa

Se propone realizar un muestreo estratificado según estas dos variables, de tamaño 600, para

1.Estimar una media o proporción

2.Comparación de medias o proporciones

3.Realizar test de independencia entre dos variables cualitativas

Objetivo 1: Estimación de una media o de una proporción

Ciertos paquetes proponen como opción una variable "peso" que permite estimar los parámetros de una población a partir de una muestra estratificada. En el ejemplo los pesos a afectar a cada individuo de cada estrato son los presentados en las Figuras 4.6 y 4.7.

	Tamaño			
Sector	- 50	50 a 100	de 100 a 500	+ 500
Industria	250	250	253	250
Comercio	251	250	244	200
Servicios	250	250	250	233

Figura 4.6 Peso para la muestra estratificada proporcional

	Tamaño			
Sector	- 50	50 a 100	de 100 a 500	+ 500
Industria	764	120	96	20
Comercio	868	80	44	8
Servicios	786	100	100	14

Figura 4.7 Peso para la muestra equi-repartida por estrato

Las medias, varianzas, covarianzas ponderadas de los valores de la muestra son estimaciones de las medias, varianzas, covarianzas de la población. Pero las estimaciones por intervalo de confianza y los tests de hipótesis *se fundamentan siempre en la hipótesis de muestreo aleatorio simple* y no de muestreo estratificado.

El estudio arrojó por resultado que la proporción de dirigentes de empresas francesas satisfechos por la decisión

gubernamental es de 0,4367. Entonces, la estimación por intervalo de confianza al 95% de la proporción p de empresarios satisfechos con la medida gubernamental:

$$0.4367 \pm 1.96 \sqrt{\frac{0.4367\ (1 - 0.4367)}{599}} = 0.4367 \pm 0.0397$$

Esto es, $[39{,}7\%; 47{,}6\%]$

Objetivo 2: Comparación de medias o de proporciones

Aquí es necesario identificar si se está en presencia de muestras independientes o apareadas. El primer caso corresponde a tener una variable observada sobre dos muestras independientes extraídas de dos subpoblaciones disjuntas; el segundo caso se corresponde con tener dos variables observadas en una misma muestra.

Si se quiere comparar las proporciones de empresarios satisfechos para las clases de tamaños 3 y 4, se está en el caso de muestras independientes, con estratificación cruzada: según el tamaño y el sector de actividad. Ambas muestras están entonces compuestas cada una de 3 muestras independientes.

Objetivo 3: Test χ^2 de independencia de dos variables cualitativas

Cuando dos variables cualitativas son las variables de estratificación (sector de actividad y tamaño, como en el ejemplo que planteado), la pregunta no tiene sentido puesto que los tamaños de las muestras fueron fijados a priori de acuerdo a la estratificación existente en la población.

Cuando una sola de las dos variables categóricas es una variable de estratificación y que es la única variable de estratificación, el test χ^2 de independencia es el test χ^2 de homogeneidad de distribuciones y puede ser utilizado para estudiar la independencia de las dos variables.

Cuando hay varias variables de estratificación y se intenta someter a una prueba la independencia deuna de ellas con una variable cualitativa que no se tomó en cuenta en la estratificación, o también cuando se quiere probar la independencia de dos variables categóricas que no se tomaron en cuenta en la estratificación, no es válido utilizar el test χ^2 de homogeneidad de distribuciones.

Ejemplo 4.3. con muestra estratificada proporcional:

Si se clasifica a los empresarios por sexo y tamaño de la empresa como se observa en la Figura 4.8. Para utilizar el test del χ^2 se reagrupan las dos últimas clases de efectivos con el fin de obtener un efectivo superior o igual a 5 en cada celda como queda expresado en la Figura 4.9.

Sexo	Tamaño				
	- 50	50 a 100	100 a 500	+ 500	Total
M	108	7	4	1	120
H	375	53	44	8	480
Total	483	60	48	9	600

Figura 4.8

	Tamaño			Total
Sexo	- 50	50 a 100	+ 100	
M	108	7	5	120
H	375	53	52	480
Total	483	60	57	600

Figura 4.9

El valor del χ^2es, con las notaciones usuales:

$$\chi^2 = \sum_{\forall i,j} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} = n\left(\sum_{\forall i,j} \frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right)$$

$$= 600\left(\frac{108^2}{433 \times 120} + \frac{7^2}{60 \times 120} + \frac{5^2}{57 \times 120} + \frac{375^2}{483 \times 480} + \frac{53^2}{60 \times 480} + \frac{52^2}{57 \times 480} - 1\right) = 8.78$$

El test de independencia del χ^2 consiste en comparar ese valor con el valor crítico de la distribución chi –cuadrado con dos grados de libertad.

Se lee en las tablas: $P(\chi^2_{2gl} < 5.99) = 0.95$

Dado que $8.78 > 5.99$, se rechaza la hipótesis de independencia de las dos variables: “sexo del empresario” y “tamaño de la empresa”.

Lamentablemente, incluso en el caso en que las variables son independientes, el valor 8.78 no puede ser considerado como la observación de una variable aleatoria según una χ^2 con dos grados de libertad, puesto que se obtienen las observaciones a partir de una muestra estratificada.

La conclusión de rechazar la hipótesis no se fundamenta más en un test de hipótesis en el sentido estadístico de la palabra. A veces se utilizan esos tests con la reserva que se acaba de mencionar.

Ejemplo 4.4. con muestra estratificada equi-repartida:

Huelga decir que en el caso de una muestra estratificada no proporcional, habrá que hacer los cálculos sobre datos ponderados.

Si la repartición según el tamaño de la empresa y el sexo del empresario es la observada en la Figura 4.10 y para una muestra de tamaño 600, después de una reponderación, es la observada en la Figura 4.11; la repartición es similar a la propuesta para la muestra estratificada proporcional y conduce entonces al mismo "test".

	Tamaño				
Sexo	- 50	50 a 100	100 a 500	+ 500	Total
M	27404	1700	1040	182	30326
H	93496	13300	10960	1918	119674
Total	120900	15000	12000	2100	150000

Figura 4.10

	Tamaño				
Sexo	- 50	50 a 100	100 a 500	+ 500	Total
M	110	7	4	1	122
H	373	53	44	8	478
Total	483	60	48	9	600

Figura 4.11

Como se dijo más arriba, una prueba de independencia permite determinar si existe asociación estadística entre dos variables categóricas. Las pruebas más utilizadas para estudiar la independencia son las chi-cuadrado de Pearson, debido a su fácil aplicación y a que se encuentran en todos los paquetes estadísticos. Estas pruebas asumen que todas las observaciones son independientes y que están igualmente distribuidas, supuestos que sólo se satisfacen para un muestreo aleatorio simple con reposición y se cumplen aproximadamente en una muestra aleatoria simple sin reposición para una fracción de muestreo pequeña.

Está bien documentado en la literatura, aunque no es conocido en la práctica, que las pruebas chi-cuadrado de Pearson dan resultados extremadamente erróneos cuando se aplican en datos obtenidos mediante diseños muestrales complejos, donde no se cumplen los supuestos, tales como el muestreo sistemático, estratificado, por

conglomerados, probabilidad variable o cualquier otro muestreo probabilístico diferente al muestreo aleatorio simple. Las implicaciones por el reporte de resultados falsos pueden ser muy costosas, dependiendo de la naturaleza del estudio que se esté realizando. Es una práctica común ignorar la complejidad del diseño muestral y proceder como si las pruebas chi-cuadrado de Pearson se comportaran de la misma forma que lo hacen bajo un muestreo aleatorio simple. De acuerdo a este criterio, se emplea un paquete estadístico estándar para aplicar las pruebas chi-cuadrado, lo que conduce a niveles de significación bastante altos.

5. ANALISIS BIVARIADO: UNA VARIABLE CUANTITATIVA Y UNA CUALITATIVA

Cuando sobre una variable cuantitativa se desea estudiar la dependencia, o efectos, que sobre su comportamiento tienen modalidades de una variable cualitativa, el método a emplear es el análisis de la varianza.

Análisis de varianza

Este método se utiliza para comparar las diferencias entre las medias de diferentes grupos, al combinar, en el mismo análisis, variables cuantitativas y variables cualitativas.

El objetivo es conocer si existen diferencias en los valores medios de la variable cuantitativa en cada modalidad de la variable cualitativa, bajo la hipótesis nula de igualdad entre las medias de los distintos grupos. Esta hipótesis se formula:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

La hipótesis alternativa es que las medias no son todas iguales

$$H_1 : \text{no todas las } \mu_j \text{ son iguales } j = 1, 2, \cdots, k$$

El estadístico de contraste tiene en cuenta las variaciones cuantitativas dentro de cada modalidad y las variaciones cuantitativas entre las modalidades. Formalmente es

$$F = \frac{MSA}{MSW} \sim F_{k-1;n-k}$$

Donde: MSA es la varianza entre grupos; MSW es la varianza dentro de los grupos; k es el número de modalidades o grupos en los que particiona la variable cualitativa y n el número total de observaciones.

Ejemplo 5.1. Se cuenta con la calificación de 15 integrantes en un programa técnico. Este programa consta de tres métodos de enseñanza -A, B y C- que desarrollan un nivel determinado de habilidad en diseño auxiliado por computadora; estos métodos fueron asignados en forma aleatoria entre los participantes.

La variable cuantitativa presente en el análisis es la calificación obtenida y la variable cualitativa el método de instrucción que consta de 3 modalidades.

La Figura 5.1 presenta las calificaciones alcanzadas, al término de la unidad de instrucción en cada método, y las calificaciones promedio correspondientes.

En primer lugar se calcula:

$$\boldsymbol{X_j = \sum_{i=1}^{n_k} X_{ij}} \qquad \forall\, \boldsymbol{j = A, B, C}$$

$$\boldsymbol{\bar{X}_j = \sum_{i=1}^{n_k} X_{ij} \Big/ n_k} \qquad \forall\, \boldsymbol{j = A, B, C}$$

Donde X_j es la suma de todas las calificaciones (en cada grupo); X_{ij} es la calificación promedio obtenida por cada participante; $\bar{X}_j$ es la media de calificaciones en cada modalidad y n_k es la cantidad de observaciones en cada grupo.

Instrucción	Calificaciones de la prueba por participante					Calificaciones Totales X_j	Calificaciones Promedio $\bar{X}_j$
	1	2	3	4	5		
A	86	79	81	70	84	400	80
B	90	76	88	82	89	425	85
C	82	68	73	71	81	375	75

Figura 5.1 Calificaciones alcanzadas por los participantes

El cálculo de la media de todos los elementos de la muestra ($\bar{X}_{..}$) se calcula a partir del doble sumatorio, uno para sumar los elementos dentro de la fila y el otro para sumar los totales entre los grupos.

$$\bar{\boldsymbol{X}}_{..} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{k}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n_k}} \boldsymbol{X_{ij}} \Big/ \boldsymbol{k} * \boldsymbol{n_k} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{3}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{5}} \boldsymbol{X_{ij}} \Big/ \boldsymbol{3 * 5} = \frac{\boldsymbol{400 + 425 + 375}}{\boldsymbol{15}} = \boldsymbol{80}$$

Para llevar a cabo una prueba de análisis de varianza (ANOVA), se subdivide la variación total (SST) en aquellas que pueden atribuirse a la variación entre grupos (SSA) y la que se debe a variaciones dentro de los grupos (SSW); de modo que

$$SST = SSA + SSW$$

Ahora se calculan estas variaciones:

variación total (SST)

$$\begin{aligned} SST &= \sum_{j=1}^{k} \sum_{i=1}^{n_k} (X_{ij} - \bar{X})^2 = \sum_{j=1}^{3} \sum_{i=1}^{5} (X_{ij} - 80)^2 \\ &= (86-80)^2 + (79-80)^2 + (81-80)^2 + (70-80)^2 + (84-80)^2 \\ &+ (90-80)^2 + (76-80)^2 + (88-80)^2 + (82-80)^2 + (89-80)^2 \\ &+ (82-80)^2 + (68-80)^2 + (73-80)^2 + (71-80)^2 + (81-80)^2 \\ &= 698 \end{aligned}$$

variación entre grupos (SSA)

$$\begin{aligned} SSA &= \sum_{j=1}^{k} n_j (\bar{X}_j - \bar{X})^2 = \sum_{j=1}^{3} 5 \left((80-80)^2 + (85-80)^2 + (75-80)^2 \right) \\ &= 5(0 + 25 + 25) = 50 * 5 = 250 \end{aligned}$$

variación dentro del grupo (SSW)

$$SSW = \sum_{j=1}^{k}\sum_{i=1}^{n_k}(X_{ij} - \bar{X}_j)^2 = \sum_{i=1}^{5}(X_{iA} - 80)^2 + \sum_{i=1}^{5}(X_{iB} - 75)^2 + \sum_{i=1}^{5}(X_{iC} - 85)^2$$
$$= (86-80)^2 + (79-80)^2 + (81-80)^2 + (70-80)^2 + (84-80)^2$$
$$+ (90-85)^2 + (76-85)^2 + (88-85)^2 + (82-85)^2 + (89-85)^2$$
$$+ (82-75)^2 + (68-75)^2 + (73-75)^2 + (71-75)^2 + (81-75)^2$$
$$= 448$$

El resultado de los cálculos anteriores se dispone en la Figura 5.6. La hipótesis nula a probar es

$$H_0: \mu_A = \mu_B = \mu_C$$

Al nivel de confianza de 0.95, el estadístico crítico es

$$F_{k-1;n-k} = F_{3-1;15-3} = F_{2;12} = 3{,}89$$

Fuente de Variación		Grados de libertad	Varianzas
Entre fila	$SSA = 250$	$k - 1 = 3 - 1 = 2$	$MSA = \frac{SSA}{k-1} = \frac{250}{2}$
Dentro de la fila	$SSW = 448$	$n - k = 15 - 3 = 12$	$MSW = \frac{SSW}{n-k} = \frac{448}{12}$
Total	$SST = 698$	$n - 1 = 15 - 1 = 14$	$MST = \frac{SST}{n-1} = \frac{698}{14}$

Figura 5.2

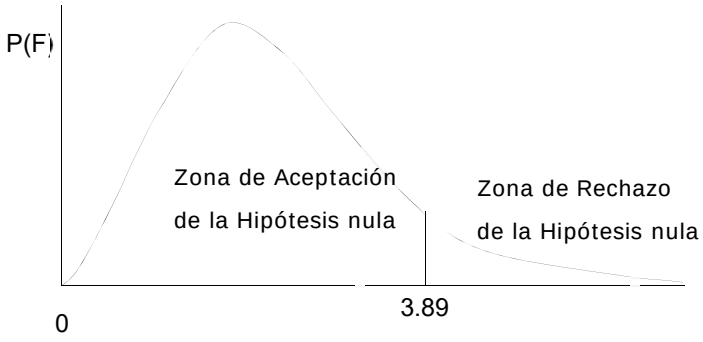


Figura 5.3 Prueba de la hipótesis de igualdad de medias

El valor empírico se calcula

$$F^e = \frac{MSA}{MSW} = 3{,}35$$

Este valor es menor que el valor crítico $\left(F^e < F_{2,12}\right)$ por lo que se acepta la hipótesis nula; las calificaciones promedios son iguales entre los diferentes métodos de instrucción.

6. ANALISIS EXPLORATORIO MULTIVARIADO

Para determinar la semejanza entre las unidades de observación se actúa de acuerdo al tipo de variables a analizar. Si las variables son cuantitativas se aplica el análisis de componentes principales; mientras que, si son cualitativas el método es el análisis factorial de correspondencias múltiples. En ambos casos, se puede realizar la clasificación y el agrupamiento de las unidades observadas de acuerdo a su semejanza.

Una vez realizados correctamente los cuatro primeros pasos del PIE, si la investigación se basa en el estudio de grandes tablas de datos con muchos individuos y muchas características, para el de Análisis de la Información se puede realizar análisis exploratorio.

Esta técnica trabaja sobre la tabla de datos de individuos por variables con el objetivo de evaluar:

1) La semejanza entre los individuos a través de los atributos seleccionados y
2) La asociación entre las variables seleccionadas y observadas sobre el conjunto de unidades de observación.

Si la tabla de datos contiene sólo variables cuantitativas, se puede realizar un análisis de componentes principales; mientras que, si las variables son cualitativas, el análisis a realizar es el factorial de correspondencias. Este puede ser análisis factorial de correspondencias simple (AFCS), o análisis factorial de correspondencias múltiples (AFCM).

Si sólo interesan dos atributos cualitativos observados -es decir en dos columnas de la tabla de códigos condensados- se hace necesario el **análisis de la información** basado en una tabla de contingencia sobre la cual se aplica el AFCS.

Si el interés se centra en el análisis de la totalidad de atributos observados se hace necesario el **análisis de la información** basado en varias tablas de contingencia sobre la cual se aplica AFCM.

6.1. Análisis factorial de correspondencias múltiples

Una tabla de códigos condensados es aquella donde las modalidades de las variables cualitativas se identifican con códigos.

El análisis factorial es una técnica que se utiliza para el estudio y la interpretación de las correlaciones existentes entre un grupo de variables, con el objeto de descubrir los posibles factores comunes a todas ellas.

El origen de estas técnicas se puede remontar a 1904, año en que el estadístico Ch.Spearman publicó el primer trabajo sobre análisis factorial, aunque limitado a la búsqueda de un único factor común entre varias variables. Su extensión a varios factores se debe a L. Thurstone, quien en 1941 publicó la obra “Factorial Studies on Inelligence”. Esta técnica suele inscribirse dentro del conjunto del análisis multivariable. De acuerdo con Blalock (1977), el fundamento del análisis factorial se encuentra en la idea de que si hay un gran número de índices o variables correlacionadas entre sí, estas relaciones pueden deberse a la presencia de una o más variables o factores subyacentes relacionados en grado diverso con aquellos. Se supone que las altas interrelaciones dentro de un grupo de variables se deben a una o varias variables o factores generales a las que representa el grupo. Precisamente, uno de los objetivos principales del análisis factorial es identificar estos factores o variables comunes y, por tanto, más generales que los datos. Con ello frecuentemente reciben significación muchos conjuntos de correlaciones que, de otra manera, parecía que carecían de sentido.

Justificación

El análisis factorial está justificado cuando, como dice Schuessler (1971), hay un conjunto de intercorrelaciones significativas entre variables de diferente ámbito. Por ejemplo, variables sociales y económicas: riqueza, empleo, desarrollo, divorcio, criminalidad, escolarización, entre otras; referidas a ciudades o núcleos de población y que da la impresión de constituir una mezcla de cifras sin un sentido coherente.

Es oportuno utilizar el análisis factorial para determinar si dicho conjunto de intercorrelaciones, aparentemente inconexas, se deben a uno o varios factores o variables no explícitas, con los cuales las variables iniciales se hallan fuertemente correlacionadas. Esta técnica se puede asimilar al método de las concordancias de Stuart Mill; de acuerdo a él, cuando se descubre un factor que es común a una serie de variables, que en otro caso no tendrían relación entre sí, tal factor se puede considerar causa común de la intercorrelación observada entre las variables dispares.

Interpretación de los factores

El análisis factorial revela estadísticamente los factores comunes más generales que unen las variables dadas en el análisis, pero no los identifica; es decir, no da su nombre, sino únicamente un resultado numérico.

Para identificar o determinar el nombre de estos factores es necesario interpretar estos resultados numéricos, lo que se realiza viendo qué variables iniciales de las correlacionadas tienen valoraciones altas en cada factor. También hay que tener en cuenta si estas valoraciones son positivas o negativas o ambas cosas a la vez.

El examen de las variables iniciales -que tienen alta valoración de las coordenadas en los factores- y del signo –que asumen estas coordenadas- debe conducir a identificar y nombrar los factores comunes, de acuerdo a las características observadas. En cierto modo, conjeturar el carácter del factor común descubierto y darle el nombre apropiado es un problema.

Ejemplo 6.1. En una investigación sobre relaciones familiares, mediante este análisis se descubre que un factor tiene correlaciones positivas con varias variables referentes a los padres y correlaciones negativas con variables relativas a los hijos, entonces a este factor se le podría llamar factor padres versus hijo.

Ejemplo 6.2. La encuesta de un diario regional.

Un diario regional realizó en 1999 una encuesta a 505 hogares en la región donde el diario tenía influencia. El objetivo general de esta encuesta era "conocer mejor a sus lectores" saber lo que se leía en el diario y responder a la pregunta: "¿Quién lee qué?".

Secciones	efectivos	peso
Noticias internacionales	128	39.75
Clasificados	175	54.35
Noticias sobre sexo, violencia, drogas	124	38.51
Noticias policiales	91	28.26
Noticias nacionales económicas	74	22.98
Noticias nacionales políticas	90	27.95
Información sobre la corrupción política y social	104	32.30
Deportes	122	37.89
Comentarios científicos	166	51.55
Información gremial	179	55.59
Notas editoriales	168	52.17
Comentarios firmados por analistas económicos y políticos	141	43.79
Comentarios firmados por analistas locales sobre aspectos regionales	103	31.99
Noticias y comentarios agropecuarios	154	47.83
Noticias regionales	52	16.15
Investigaciones periodísticas	126	39.13
Información sobre espectáculos	145	45.03
Sociales	121	37.58
Chistes, horóscopo y entretenimiento en general	123	38.20
Suplemento y revista dominical	117	36.34

Figura 6.1 Secciones del diario regional (efectivos y peso de sus lectores)

Para ello fueron seleccionadas 20 secciones que aparecían diariamente y que cubrían, aproximadamente, la totalidad de los temas abordados en el diario, las que se encuentran indicadas en la Figura 6.1.

A cada encuestado se lo consultó, en primer término, si leía el diario. Si la respuesta era positiva entonces se le pidió que indicara que actitud tenía hacia ella: si la leía, si no la leía, o si no la conocía

(Ns/Nc). Esto da lugar a que, para cada sección existan tres respuestas posibles. Las secciones no son mutuamente excluyentes entre sí por lo que cada sección es una variable en sí misma; las respuestas alternativas en cada sección sí son mutuamente excluyentes y cada una de estas respuestas es una modalidad de las variables cualitativas a la que pertenece.

Además del tema particular de investigación, las encuestas contienen un cuestionario relativo a la caracterización del encuestado, Figura 6.2. Esto es así porque cada encuestado no interesa en sí mismo sino como representante de ciertas categorías de población.

Cómo realizar el Análisis

Generalmente, las tablas resultantes de un relevamiento no constituyen un conjunto homogéneo, en el sentido que contienen información sobre diferentes ámbitos de la población bajo estudio; es decir, están estructuradas en varios grupos. En las encuestas que relevan información de hogares o de personas, el tema objeto de investigación siempre va acompañado de información que permita caracterizar al entrevistado desde el punto de vista socioeconómico porque esto va a permitir luego la segmentación de la población.

Si el estudio está relacionado con la penetración de un producto en el mercado, la segmentación de la que se habla es la del mercado para que el empresario sepa cuál es su porción en él y las características de sus consumidores.

Variable	Modalidad	Efectivos	Peso
Sexo	Hombre	162	50.31
	Mujer	160	49.69
Es jefe de familia?	SI	152	47.20
	NO	170	52.80
Actividad que desarrolla	Desocupado	8	2.48
	Estudiante	28	8.70
	Ama de casa	55	17.08
	Jubilado	39	12.11
	Empleadores	10	3.11
	Trabajo no especializado o changarín	14	4.35
	Comerciante sin personal	29	9.01
	Técnico	13	4.04
	Profesional independiente y Otros	25	7.76
	autónomos	12	3.73
	Relación de dependencia	12	3.73
	Obrero, calificado y no calificado,	67	20.81
	capataz	10	3.11
	Empleado sin jerarquía		
	Jefe intermedio		
Nivel de Educación	Primario incompleto	23	7.14
	Primario completo	89	27.64
	Secundario incompleto	67	20.81
	Secundario completo	77	23.91
	Terciario incompleto	13	4.04
	Terciario completo	25	7.76
	Universitario incompleto	17	5.28
	Universitario completo	5	1.55
	Ns/Nc	6	1.86
Nivel de Ingresos	Menos de $300.00	49	15.22
	De $300.00 a $500.00	95	29.50
	De $501.00 a $700.00	68	21.12
	De $701.00 a $1000.00	45	13.98
	De $1001.00 a $1500.00	28	8.70
	De $1501.00 a $2500.00	19	5.90
	De $2501.00 a $4000.00	6	1.86
	Más de $4000.00	2	0.62
	Ns/Nc	10	3.11
Número de personas en el hogar	Una persona	10	3.11
	Dos personas	40	12.42
	De 3 a 4 personas	145	45.03
	De 5 a 6 personas	84	26.09
	Más de 7 personas	43	13.35
Tiene Computadora	Si	42	13.04
	No	280	86.96
Tiene Televisor	Si	316	98.14
	No	6	1.86
Tiene horno a microondas	Si	41	12.73
	No	281	87.27
Tiene radio	Si	315	97.83
	No	7	2.17
Tiene auto	Si	151	46.89
	No	171	53.11
Tiene video	Si	102	31.68
	No	220	68.32
Nivel de la vivienda	Media superior	29	9.01
	Media-Media	114	35.40
	Media Baja	118	36.65
	Muy modesta o Marginal	61	18.94

Figura 6.2 La caracterización de los efectivos y pesos de sus modalidades

Un AFCM para toda la base, sin selección de un grupo en particular, es un análisis simultáneo de las dimensiones de análisis y dan por resultado la respuesta a una pregunta compleja. Pero también es posible analizar los diferentes ámbitos por separado; en cuyo caso se hace uso de los elementos suplementarios. Esto consiste en dejar como activas todas las variables que se corresponden con un ámbito del estudio y a las restantes incorporarlas como suplementarias o ilustrativas.

El Análisis Factorial Múltiple, al analizar tablas compuestas de varios grupos de variables, aporta una solución más satisfactoria al problema del equilibrio de los grupos que otros análisis.

En la encuesta del diario regional que contiene dos ámbitos de variables cualitativas, son posibles tres soluciones:

- Un AFCM con el conjunto de las secciones y de la caracterización como principal, que es el análisis que brindará la respuesta a la pregunta “¿quién lee qué?”
- Un AFCM del conjunto de las secciones (como principal) y de la caracterización (como suplementario), que es el análisis que brindará la respuesta a la pregunta “¿qué se lee?”
- Un AFCM del conjunto de la caracterización (como principal) y de las secciones (como suplementario), que es el análisis que brindará la respuesta a la pregunta “¿quién lee?”

Para responder a la pregunta “¿quién lee qué?”, el AFCM encontrará

- Una tipología de los individuos según su perfil de lectura y su caracterización: dos individuos son próximos si se interesan o no se interesan por las mismas secciones del diario y si tienen similares características socioeconómicas.
- Un estudio de la relación entre las variables activas y suplementarias siempre que se decide darle esta categoría a alguna de las variables activas.

El AFCM que permite responder la pregunta “¿qué se lee?” permitirá encontrar

- Una tipología de los individuos según su perfil de lectura: dos individuos son próximos si se interesan o no se interesan por las mismas secciones del diario
- Un estudio de las relaciones entre la lectura (o la no lectura) de las diferentes secciones. Si varias secciones son, a menudo, leídas por los mismos lectores constituyen un grupo que será puesto en evidencia; también, serán detectados los fenómenos de exclusión (los lectores de la sección A no leen la sección B).
- Un estudio de la relación entre los principales factores de variabilidad de los perfiles lectura con cada una de las variables de caracterización (elementos suplementarios), tomadas por separado.

Responder a la pregunta “¿quién lee?” es la solución inversa a la precedente. Las modalidades de las 7 variables de caracterización son los elementos principales y las modalidades perfiles de lectura son los elementos suplementarios. Del AFCM de las variables de caracterización puede esperarse:

- Una tipología de los individuos según su caracterización: dos individuos son próximos si sus características se parecen independientemente de su lectura
- Un estudio de las relaciones entre las diferentes variables de la caracterización
- Un estudio de la relación entre el perfil característico general y la lectura de cada sección como elemento suplementario.

Lectura de la tabla de datos

Las modalidades utilizadas en las variables cualitativas son categorías mutuamente excluyentes entre sí. De modo que la j-ésima característica observada está compuesta de k_j modalidades mutuamente excluyentes. Por ejemplo, la variable Información internacional se compone de las modalidades: lee la sección, no lee la sección pero la conoce y no conoce la sección.

Si se observa la Figura 6.3 se encontrará el número efectivo de individuos (eff) que presenta una modalidad determinada dentro de cada variable considerada. Este número de individuos son los puntos (poids) considerados para el análisis. Cuando se observa la j-ésima característica sobre el i-ésimo individuo, se puede afectar la k-ésima modalidad de esa j-ésima característica del i-ésimo individuo.

En la Figura 6.4, por ejemplo, el individuo 7 asume la modalidad 2 en noticias policiales, esto equivale a decir “no leo la sección pero la conozco”. La Figura 6.4 ilustra una tabla de códigos condensados que se construye a partir de la codificación de la información.

```
ANALYSE DES CORRESPONDANCES MULTIPLES
APUREMENT DES MODALITES ACTIVES
SEUIL (PCMIN)   :      2.00 %       POIDS:       6.44
AVANT APUREMENT :     20 QUESTIONS ACTIVES         60 MODALITES ASSOCIEES
APRES           :     20 QUESTIONS ACTIVES         60 MODALITES ASSOCIEES
POIDS TOTAL DES INDIVIDUS ACTIFS :    322.00
TRI-A-PLAT DES QUESTIONS ACTIVES
----------------------------+----------------+-----------------------------------------------------------------------
         MODALITES          | AVANT APUREMENT |           APRES APUREMENT
IDENT      LIBELLE          |  EFF.     POIDS |  EFF.     POIDS       HISTOGRAMME DES POIDS RELATIFS
----------------------------+----------------+-----------------------------------------------------------------------
    3 . Noticias Internacionales
Int1 - Int1                 |  128     128.00 |  128     128.00  ************************
Int2 - Int2                 |  167     167.00 |  167     167.00  *********************************
Int3 - Int3                 |   27      27.00 |   27      27.00  ******
----------------------------+----------------+-----------------------------------------------------------------------
    4 . Clasificados
Cla1 - Cla1                 |  175     175.00 |  175     175.00  *********************************
Cla2 - Cla2                 |   91      91.00 |   91      91.00  *****************
Cla3 - Cla3                 |   56      56.00 |   56      56.00  ***********
----------------------------+----------------+-----------------------------------------------------------------------
    5 . Noticias de Violencia, sexo, droga
Vio1 - Vio1                 |  124     124.00 |  124     124.00  ************************
Vio2 - Vio2                 |  146     146.00 |  146     146.00  ****************************
Vio3 - Vio3                 |   52      52.00 |   52      52.00  **********
----------------------------+----------------+-----------------------------------------------------------------------
    6 . Noticias policiales
Pol1 - Pol1                 |   91      91.00 |   91      91.00  *****************
Pol2 - Pol2                 |  208     208.00 |  208     208.00  ****************************************
Pol3 - Pol3                 |   23      23.00 |   23      23.00  *****
----------------------------+----------------+-----------------------------------------------------------------------
    7 . Noticias nacionales económicas
Nae1 - Nae1                 |   74      74.00 |   74      74.00  **************
Nae2 - Nae2                 |  228     228.00 |  228     228.00  ********************************************
Nae3 - Nae3                 |   20      20.00 |   20      20.00  ****
----------------------------+----------------+-----------------------------------------------------------------------
    8 . Noticias nacionales políticas
Nap1 - Nap1                 |   90      90.00 |   90      90.00  *****************
Nap2 - Nap2                 |  207     207.00 |  207     207.00  ****************************************
Nap3 - Nap3                 |   25      25.00 |   25      25.00  *****
----------------------------+----------------+-----------------------------------------------------------------------
```

Figura 12.3. Continúa…

Figura 12.3. Continuación

```
    9 . Información sobre corrupción política y social
Cor1 - Cor1                 |  104     104.00 |  104     104.00  ********************
Cor2 - Cor2                 |  172     172.00 |  172     172.00  *********************************
Cor3 - Cor3                 |   46      46.00 |   46      46.00  *********
----------------------------+----------------+-----------------------------------------------------------------
   10 . Deportes
Dep1 - Dep1                 |  122     122.00 |  122     122.00  **********************
Dep2 - Dep2                 |  174     174.00 |  174     174.00  *********************************
Dep3 - Dep3                 |   26      26.00 |   26      26.00  *****
----------------------------+----------------+-----------------------------------------------------------------
   11 . Comentarios Científicos
Coc1 - Coc1                 |  166     166.00 |  166     166.00  ******************************
Coc2 - Coc2                 |   97      97.00 |   97      97.00  ******************
Coc3 - Coc3                 |   59      59.00 |   59      59.00  ***********
----------------------------+----------------+-----------------------------------------------------------------
   12 . Información  Gremial
Gre1 - Gre1                 |  179     179.00 |  179     179.00  **********************************
Gre2 - Gre2                 |   69      69.00 |   69      69.00  *************
Gre3 - Gre3                 |   74      74.00 |   74      74.00  **************
----------------------------+----------------+-----------------------------------------------------------------
   13 . Notas Editoriales
Edi1 - Edi1                 |  168     168.00 |  168     168.00  ******************************
Edi2 - Edi2                 |   85      85.00 |   85      85.00  ****************
Edi3 - Edi3                 |   69      69.00 |   69      69.00  *************
----------------------------+----------------+-----------------------------------------------------------------
   14 . Comentarios económicos y políticos
Coe1 - Coe1                 |  141     141.00 |  141     141.00  ***************************
Coe2 - Coe2                 |  129     129.00 |  129     129.00  ************************
Coe3 - Coe3                 |   52      52.00 |   52      52.00  **********
----------------------------+----------------+-----------------------------------------------------------------
   15 . Comentarios regionales
Cor1 - Cr1                  |  103     103.00 |  103     103.00  ********************
Cor2 - Cr2                  |  162     162.00 |  162     162.00  ******************************
Cor3 - Cr3                  |   57      57.00 |   57      57.00  ***********
----------------------------+----------------+-----------------------------------------------------------------
   16 . Noticias Agropecuarias
Agr1 - Agr1                 |  154     154.00 |  154     154.00  ****************************
Agr2 - Agr2                 |  106     106.00 |  106     106.00  *******************
Agr3 - Agr3                 |   62      62.00 |   62      62.00  ************
----------------------------+----------------+-----------------------------------------------------------------
   17 . Noticias Regionales
Reg1 - Reg1                 |   52      52.00 |   52      52.00  **********
Reg2 - Reg2                 |  221     221.00 |  221     221.00  ****************************************
Reg3 - Reg3                 |   49      49.00 |   49      49.00  **********
----------------------------+----------------+-----------------------------------------------------------------
   18 . Investigación periodística
Inv1 - Inv1                 |  126     126.00 |  126     126.00  ************************
Inv2 - Inv2                 |  138     138.00 |  138     138.00  *************************
Inv3 - Inv3                 |   58      58.00 |   58      58.00  ***********
----------------------------+----------------+-----------------------------------------------------------------
   19 . Información sobre Espectáculos
Esp1 - Esp1                 |  145     145.00 |  145     145.00  ****************************
Esp2 - Esp2                 |  123     123.00 |  123     123.00  ***********************
Esp3 - Esp3                 |   54      54.00 |   54      54.00  ***********
----------------------------+----------------+-----------------------------------------------------------------
   20 . Sociales
Soc1 - Soc1                 |  121     121.00 |  121     121.00  **********************
Soc2 - Soc2                 |  141     141.00 |  141     141.00  *************************
Soc3 - Soc3                 |   60      60.00 |   60      60.00  ************
----------------------------+----------------+-----------------------------------------------------------------
   21 . Chistes, Horoscopos y entretenimiento
Chi1 - Chi1                 |  123     123.00 |  123     123.00  **********************
Chi2 - Chi2                 |  166     166.00 |  166     166.00  ******************************
Chi3 - Chi3                 |   33      33.00 |   33      33.00  *******
----------------------------+----------------+-----------------------------------------------------------------
   22 . Suplemento y revista dominical
Sup1 - Sup1                 |  117     117.00 |  117     117.00  ********************
Sup2 - Sup2                 |  153     153.00 |  153     153.00  ****************************
Sup3 - Sup3                 |   52      52.00 |   52      52.00  **********
----------------------------+----------------+-----------------------------------------------------------------
```

Figura 6.3. Variables activas y sus modalidades

Cada línea contiene los códigos correspondientes a las modalidades atribuidas a un individuo para cada una de las características observadas. Los valores que figuran en la intersección de los individuos con las modalidades de las variables son códigos y, precisamente por tratarse de una tabla de códigos, carece de propiedades numéricas. Esta tabla tiene tantas filas como individuos se consideren y tantas

columnas como la suma de modalidades contempladas en todas las variables que forman parte del análisis.

Individuo	Características							
	Noticia in-ternacional	Clasificados	Noticias sobre sexo, drogas, violencia	Noticias policial	Noticias nacionales económic	Noticias nacionales políticas	.....	Nivel de vivienda
1	2	2	2	2	2	2		4
2	1	1	1	2	2	2		4
4	2	2	2	2	2	2		4
5	2	2	2	2	2	2		4
6	2	1	1	1	2	2		4
7	2	2	2	2	2	2		4
10	2	1	1	1	2	2		3
11	2	2	2	2	2	2		3
12	1	1	1	1	1	1		4
....								
....								
....								
....								
....								
....								
503	2	1	2	2	2	2	.....	2

Figura 6.4. Tabla de datos o Tabla de códigos condensados (nxk)

El análisis factorial de correspondencias de esta tabla de códigos tiene por objetivo

- Hallar semejanza de individuos
- Estudiar relación entre variables
- Resumir las características observadas en un pequeño número de variables
- Comparar modalidades de diferentes variables

En el proceso de análisis, esta tabla de códigos condensados se convierte en una tabla lógica (Figura 6.5), donde la existencia o ausencia de una modalidad se denota con 1 o 0.

Observa-ciones	Características									
	Noticias Internacionales			Clasificados...				Nivel de vivienda		
Individuo	Lee	No lee	No conoce	Lee	No lee	No conoce	....	Media Media	Media Baja	Muy modesta o marginal
1	0	1	0	0	1	0		0	0	1
2	1	0	0	1	0	0		0	0	1
4	0	1	0	0	1	0		0	0	1
5	0	1	0	0	1	0		0	0	1
6	0	1	0	1	0	0		0	0	1
7	0	1	0	0	1	0		0	0	1
10	0	1	0	1	0	0		0	1	0
11	0	1	0	0	1	0		0	1	0
12	1	0	0	1	0	0		0	0	1
....										
....										
....										
....										
....										
....										
503	0	1	0	1	0	0		1	0	0

Figura 6.5. Tabla lógica (nxp)

Individuos

Esta tabla de códigos se representa en un espacio de np dimensiones (n individuos y p características observadas) donde las coordenadas del individuo i-ésimo viene dada por

$$\frac{x_{ij}}{p\sqrt{\frac{n_j}{np}}}$$

donde x_{ij} representa el individuo i asociado a una modalidad j de una de las p características posibles, asume el valor 1 si la modalidad se presenta y el valor cero si no se presenta

p es el número de características medidas

n_j es el número de individuos que asumen la característica j

np es el número de individuos por las p características

Es importante conocer las coordenadas de los individuos porque esto conduce a conocer la distancia entre ambos. Si dos individuos son semejantes (o no), de modo que se pueda afirmar que pertenecen a un mismo grupo de individuos (o no), depende de su distancia; es decir, de qué tan cerca se encuentren en el espacio.

Esta distancia se conoce como distancia del Chi^2 y se define como

$$d^2_{(i,i')} = \frac{1}{p}\sum_{j=1}^{K}\frac{n}{n_j}(x_{ij} - x_{i'j})^2$$

En resumen:

- cada término de la sumatoria $(x_{ij} - x_{i'j})^2$ vale 1 o 0. Vale 1 si dos individuos considerados no presentan simultáneamente la misma modalidad y vale 0 si se presenta la misma modalidad
- las distancias entre los individuos crece a medida que aumentan las diferencias de modalidades
- cada modalidad interviene en el cálculo de la distancia con la inversa de su peso; es decir $\frac{n}{n_j}$

la distancia del Chi^2 respeta el criterio de comparación adoptado

Modalidades

De igual manera que con los individuos se debe proceder a conocer la distancia entre modalidades. La coordenada de la modalidad j se define como

$$\frac{\frac{x_{nj}}{np}}{\frac{n_j}{np}\sqrt{\frac{p}{np}}}$$

donde x_{nj}representa la existencia de un individuo con esa modalidad, puede valer 1 o 0 de acuerdo a que la modalidad se presente o no

np es el número de individuos por las p características

p es el número de características medidas

n_j es el número de individuos que asumen la característica j

Un nuevo referencial de representación

Ahora bien, se tiene información de los individuos y las modalidades en un espacio de np dimensiones. Esto no puede observarse gráficamente y es necesario, para continuar con el análisis, referirlo a otro sistema de coordenadas de manera tal que el primer plano (las coordenadas de los dos primeros ejes) brinde la mejor representación de la tabla de datos. Para esto es necesario en primer término conocer el centro de gravedad de cada nube de puntos.

La nube de puntos individuos se representan en el espacio de las modalidades y tiene un centro de gravedad cuyas coordenadas son

$$\left[\sqrt{\frac{n_1}{np}};\ldots;\sqrt{\frac{n_j}{np}};\ldots;\sqrt{\frac{n_K}{np}}\right] \quad \forall j=1,\ldots,K\,.$$

La nube de puntos modalidades se representa en el espacio de los individuos y tiene un centro de gravedad cuyas coordenadas vienen dadas por la raíz cuadrada de la inversa del número total de individuos, de modo que se cuenta con un eje por cada individuo. Las coordenadas son

$$\left[\sqrt{\frac{1}{n}};\ldots;\sqrt{\frac{1}{n}};\ldots;\sqrt{\frac{1}{n}}\right] \quad \forall i=1,\ldots,n\,,$$

Inercia

Una vez conocido el centro de gravedad es necesario referir las coordenadas de todos los individuos y de todas las modalidades con su respectivo centro de gravedad, dando lugar a las denominadas matrices de inercia.

La inercia mide la variabilidad o dispersión de una nube de puntos respecto de un punto en común: el baricentro o centro de gravedad de la nube. La inercia de la nube de puntos, respecto al baricentro, se define como la suma del producto del cuadrado de la distancia de todos los puntos respecto del baricentro por el peso asociado a cada punto. En términos algebraicos se define como

$$I_{GC}=\sum_{j=1}^{K} p_j d^2_{(j,G_C)}=\sum_{j=1}^{K}\frac{n_j}{np}d^2_{(j,G_C)}=\left[\sum_{j=1}^{K}\frac{1}{p}\left(1-\frac{n_j}{n}\right)\right]$$

Esto significa que si la modalidad es rara, de modo que n_j tienda a 0, la contribución a la inercia va a ser importante. La contribución de cada modalidad se define como

$$Contr_j\ a\ I_{GC} = \frac{1}{p}\left(1 - \frac{n_j}{n}\right)$$

En definitiva la inercia no es otra cosa que la suma de las contribuciones de cada punto respecto del baricentro, ponderado por su peso en la nube de puntos.

La inercia global de la nube de puntos de modalidades respecto de G_C será

$$\sum_{j=1}^{K} Contr_j\ a\ I_{G_C} = \left[\sum_{j=1}^{K} \frac{1}{p}\left(1 - \frac{n_j}{n}\right)\right] = \sum_{j=1}^{K} \frac{1}{p} - \sum_{j=1}^{K} \frac{n_j}{np}$$

Es decir que

$$I_{G_C}^{N_j} = \frac{K}{p} - 1$$

De modo que la inercia total dependerá del número de características observadas (p) y del número total de modalidades (K) que presentaron esas características

Anteriormente se habían dicho dos cosas:

- que se necesitaba un nuevo referencial para resumir la información y proyectarla en un espacio de dos dimensiones
- que el nuevo referencial se halla a partir de las matrices de inercia

La matriz de inercia se obtiene al centrar todas las coordenadas de las modalidades; es decir, obtener la distancia respecto del baricentro. Al diagonalizar la matriz de inercia se encuentran los valores propios de la matriz; estos valores propios dan origen a las direcciones principales de la nube de puntos. Es indistinto hallar las direcciones principales de la matriz de inercia de las modalidades o de los individuos; en general,

se trabaja con la de modalidades porque es un espacio de menores dimensiones, habitualmente la cantidad de individuos es mayor que la cantidad de modalidades.

La inercia total de la nube de puntos es la diferencia entre el cociente de la cantidad total de modalidades que forman parte del análisis (K) y la cantidad de variables consideradas (p) y 1. De modo que

$$I_{G_C}^{N_j} = \frac{K}{p} - 1$$

Ejemplo 6.3. En el estudio del diario regional, el espacio individuos tiene 322 dimensiones; es decir, una por cada individuo

$$I_{G_C}^{N_j} = \frac{K}{p} - 1 = \frac{60}{20} - 1 = 2$$

En la Figura 6.6 se observa que la traza de la matriz de inercia, antes de la diagonalización, es igual a la suma de los valores propios: 2; siendo la suma de los valores propios la inercia total de la nube. Precisamente, descomponer la inercia significa encontrar los valores propios; el valor asumido por el valor propio es la inercia contenida en ese eje.

El método de diagonalización asegura que el primer valor propio es el más importante, por ende el factor asociado a él será la dirección principal de la nube de puntos.

La tabla informa:

- el valor asumido por el primer valor propio, 0.6276
- el peso que tiene en el total, 31.38
- el porcentaje acumulado para los dos primeros ejes, 49.44

Al incorporar un eje adicional en el análisis, la información puede representarse en planos, con los dos primeros valores propios se

obtiene el primer plano factorial donde se representa el 49.44% del total de la inercia. En la Figura 6.6 se muestra el histograma de los 40 primeros valores propios, para este último se alcanza la descomposición del 100% de la información.

APERCU DE LA PRECISION DES CALCULS : TRACE AVANT DIAGONALISATION .. 2.0000
SOMME DES VALEURS PROPRES 2.0000
HISTOGRAMME DES 40 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	0.6276	31.38	31.38	**********
2	0.3612	18.06	49.44	**********
3	0.0914	4.57	54.01	**********
4	0.0759	3.80	57.80	**********
5	0.0716	3.58	61.38	**********
6	0.0622	3.11	64.49	********
7	0.0520	2.60	67.09	*******
8	0.0476	2.38	69.47	*******
9	0.0432	2.16	71.63	******
10	0.0395	1.98	73.61	******
11	0.0365	1.82	75.43	*****
12	0.0353	1.76	77.20	*****
13	0.0333	1.67	78.86	*****
14	0.0310	1.55	80.41	****
15	0.0300	1.50	81.91	****
16	0.0286	1.43	83.34	****
17	0.0266	1.33	84.67	****
18	0.0238	1.19	85.86	****
19	0.0232	1.16	87.02	***
20	0.0222	1.11	88.13	***
21	0.0216	1.08	89.21	***
22	0.0197	0.99	90.20	***
23	0.0182	0.91	91.11	***
24	0.0178	0.89	92.00	***
25	0.0175	0.87	92.87	***
26	0.0151	0.75	93.63	**
27	0.0144	0.72	94.35	**
28	0.0136	0.68	95.03	**
29	0.0134	0.67	95.70	**
30	0.0127	0.63	96.33	**
31	0.0112	0.56	96.89	**
32	0.0100	0.50	97.39	**
33	0.0098	0.49	97.88	**
34	0.0088	0.44	98.32	**
35	0.0078	0.39	98.71	*
36	0.0077	0.38	99.09	*
37	0.0055	0.27	99.37	*
38	0.0049	0.24	99.61	*
39	0.0046	0.23	99.84	*
40	0.0032	0.16	100.00	*

Figura 6.6. Valores propios de la matriz de inercia

La cantidad de ejes a conservar en el análisis viene dada por el cociente entre la inercia total y el total de valores propios. El resultado de ese cociente se compara con los valores de los autovalores y deberán conservarse todos los ejes que se encuentren por encima de él. En este caso los 7 primeros valores propios. Ahora bien, la información que se agrega a partir del tercer eje puede ser despreciable; en este caso particular, el

aporte del tercer eje es significativamente menor. En general se analiza el plano y puede en algún caso estudiarse el tercer eje.

El nuevo referencial presenta tres ventajas importantes

- Permite analizar completamente la forma de la nube de puntos perfiles
- Permite la representación de la forma de la nube de puntos perfiles cualquiera sea la dimensión de la nube de puntos
- Permite producir una representación objetiva de esas nubes de puntos independientes del punto de vista del análisis

b. El plano de los dos primeros factores

Las variables activas

Los puntos en el espacio que representan las modalidades de las variables utilizadas para el análisis deben referirse al nuevo sistema de coordenadas. Esto significa que deben proyectarse sobre las direcciones principales halladas al descomponer la inercia.

Cada modalidad tiene un peso relativo en la nube de puntos; este peso viene dado por la cantidad de individuos que se identifican con esa modalidad respecto del producto del total de observaciones y la cantidad de variables que componen la nube de puntos. El peso relativo de una modalidad es

$$\frac{\sum_{i=1}^{n} x_{ij}}{np} * 100$$

La distancia está referida al centro de gravedad y se obtiene haciendo

$$d^2_{(j,Gc)} = \frac{n}{n_j} - 1$$

Las coordenadas indican la representación de cada modalidad sobre el nuevo sistema de referencia. El valor de las coordenadas indica la distancia al origen del sistema y la contribución a la formación del eje; a mayor valor de coordenadas sobre un eje, mayor distancia al origen del sistema y cuánto más cerca se encuentre del eje factorial más contribuirá a su formación.

La contribución de la variable a la formación del factor surge de la suma de la contribución individual de cada modalidad perteneciente a la variable. Luego, la suma de todas las contribuciones acumuladas de las variables consideradas es igual a 100, es decir, el 100%.

El coseno cuadrado mide la calidad de representación de la distancia del punto i al origen del eje factorial. Medir la calidad de representación significa conocer si el punto individuo o modalidad está bien representado. El coseno cuadrado varía entre 0 y 1, cuanto mayor sea el coseno cuadrado mejor representada la modalidad en ese eje.

Ejemplo 6.4. Teniendo de referencia a las Figuras 6.3 y 6.7, en el estudio del diario regional la modalidad *lee* noticias internacionales tiene el peso relativo de

$$\frac{\sum_{i=1}^{n} x_{ij}}{np} * 100 = \frac{128}{322 * 20} * 100 = 1.99$$

su distancia al centro de gravedad se obtiene haciendo

$$d^2_{(j,Gc)} = \frac{n}{n_j} - 1 = \frac{322}{128} - 1 = 1.52$$

La modalidad *no lee* de noticias internacionales tiene sobre el primer eje factorial coordenada igual a –0.09, la que está muy cercana al origen del sistema y no contribuye a la formación del eje; mientras que, la modalidad *no conoce* la sección de noticias

internacionales tiene coordenada –2.07 y contribución de 2.9, ésta es muy alta si se tiene en cuenta que la variable noticias internacionales contribuye en el 3.8% a la formación del factor 1.

La contribución de la modalidad *no conoce* la sección de noticias internacionales, se calcula haciendo

$$\frac{coordenada^2 * peso}{autovalor} = \frac{-2.07^2 * 0.42}{0.6276} = 2.9$$

y su coseno cuadrado es

$$\frac{coordenadd^2}{d^2_{(j,Gc)}} = \frac{-2.07^2}{10.93} = 0.39$$

Con estos elementos se analiza la Figura 6.7

En primer lugar, se buscan aquellas variables que tengan contribución acumulada más alta en cada eje, porque esto significa que son las que más aportan a su construcción.

MODALITES			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDEN - LIBELLE	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
3 . Noticias Internacionales																	
Int1 - Int1	1.99	1.52	0.55	0.52	-0.29	-0.07	-0.20	1.0	1.5	1.8	0.1	1.1	0.20	0.18	0.05	0.00	0.03
Int2 - Int2	2.59	0.93	-0.09	-0.48	0.29	0.19	0.18	0.0	1.6	2.4	1.2	1.2	0.01	0.25	0.09	0.04	0.04
Int3 - Int3	0.42	10.93	-2.07	0.48	-0.45	-0.83	-0.19	2.9	0.3	0.9	3.8	0.2	0.39	0.02	0.02	0.06	0.00
							CONTRIBUTION CUMULEE =	3.8	3.4	5.1	5.2	2.5					
4 . Clasificados																	
Cla1 - Cla1	2.72	0.84	0.51	0.29	0.04	0.00	0.09	1.1	0.6	0.0	0.0	0.3	0.31	0.10	0.00	0.00	0.01
Cla2 - Cla2	1.41	2.54	0.13	-0.79	-0.01	-0.07	-0.24	0.0	2.4	0.0	0.1	1.2	0.01	0.25	0.00	0.00	0.02
Cla3 - Cla3	0.87	4.75	-1.81	0.39	-0.10	0.10	0.11	4.5	0.4	0.1	0.1	0.1	0.69	0.03	0.00	0.00	0.00
							CONTRIBUTION CUMULEE =	5.7	3.4	0.1	0.2	1.6					
5 . Noticias de Violencia, sexo, droga																	
Vio1 - Vio1	1.93	1.60	0.59	0.46	-0.24	0.17	0.78	1.1	1.1	1.2	0.7	16.2	0.22	0.13	0.03	0.02	0.38
Vio2 - Vio2	2.27	1.21	0.14	-0.52	0.23	-0.16	-0.70	0.1	1.7	1.3	0.8	15.3	0.02	0.22	0.04	0.02	0.40
Vio3 - Vio3	0.81	5.19	-1.80	0.36	-0.08	0.06	0.10	4.2	0.3	0.1	0.0	0.1	0.63	0.03	0.00	0.00	0.00
							CONTRIBUTION CUMULEE =	5.3	3.1	2.5	1.5	31.7					
6 . Noticias policiales																	
Pol1 - Pol1	1.41	2.54	0.56	0.54	-0.15	0.18	1.13	0.7	1.1	0.4	0.6	25.0	0.12	0.11	0.01	0.01	0.50
Pol2 - Pol2	3.23	0.55	-0.01	-0.29	0.15	0.05	-0.56	0.0	0.8	0.8	0.1	13.9	0.00	0.15	0.04	0.01	0.56
Pol3 - Pol3	0.36	13.00	-2.08	0.50	-0.74	-1.18	0.57	2.5	0.2	2.1	6.6	1.6	0.33	0.02	0.04	0.11	0.02
							CONTRIBUTION CUMULEE =	3.2	2.1	3.3	7.3	40.6					
7 . Noticias nacionales económicas																	
Nae1 - Nae1	1.15	3.35	0.56	0.90	-0.93	0.66	-0.23	0.6	2.6	10.8	6.5	0.8	0.09	0.24	0.26	0.13	0.02
Nae2 - Nae2	3.54	0.41	0.00	-0.34	0.44	-0.01	0.07	0.0	1.1	7.4	0.0	0.2	0.00	0.27	0.46	0.00	0.01
Nae3 - Nae3	0.31	15.10	-2.07	0.50	-1.53	-2.32	0.08	2.1	0.2	8.0	22.1	0.0	0.28	0.02	0.16	0.36	0.00
							CONTRIBUTION CUMULEE =	2.7	3.9	26.2	28.6	1.1					
8 . Noticias nacionales políticas																	
Nap1 - Nap1	1.40	2.58	0.59	0.75	-0.77	0.55	-0.39	0.8	2.1	9.0	5.7	2.9	0.13	0.22	0.23	0.12	0.06
Nap2 - Nap2	3.21	0.56	-0.02	-0.39	0.47	-0.05	0.18	0.0	1.4	7.7	0.1	1.5	0.00	0.27	0.40	0.01	0.06
Nap3 - Nap3	0.39	11.88	-1.93	0.55	-1.12	-1.55	-0.10	2.3	0.3	5.3	12.3	0.1	0.31	0.03	0.11	0.20	0.00
							CONTRIBUTION CUMULEE =	3.1	3.8	22.1	18.1	4.4					
9 . Información sobre corrupción política y social																	
Cor1 - Cor1	1.61	2.10	0.65	0.69	-0.38	0.30	0.00	1.1	2.2	2.5	1.9	0.0	0.20	0.23	0.07	0.04	0.00
Cor2 - Cor2	2.67	0.87	0.09	-0.55	0.28	-0.21	-0.01	0.0	2.2	2.3	1.6	0.0	0.01	0.34	0.09	0.05	0.00
Cor3 - Cor3	0.71	6.00	-1.83	0.47	-0.20	0.11	0.05	3.8	0.4	0.3	0.1	0.0	0.56	0.04	0.01	0.00	0.00
							CONTRIBUTION CUMULEE =	5.0	4.8	5.2	3.7	0.0					
10 . Deportes																	
Dep1 - Dep1	1.89	1.64	0.47	0.30	-0.07	0.24	-0.16	0.7	0.5	0.1	1.5	0.7	0.13	0.05	0.00	0.04	0.02
Dep2 - Dep2	2.70	0.85	-0.04	-0.29	0.12	0.02	0.12	0.0	0.6	0.4	0.0	0.6	0.00	0.10	0.02	0.00	0.02
Dep3 - Dep3	0.40	11.38	-1.90	0.50	-0.47	-1.24	-0.07	2.3	0.3	1.0	8.2	0.0	0.32	0.02	0.02	0.14	0.00
							CONTRIBUTION CUMULEE =	3.0	1.4	1.5	9.6	1.3					
11 . Comentarios Científicos																	

Coc1 - Coc1	2.58	0.94	0.62	0.48	0.06	-0.07	-0.13	1.6	1.6	0.1	0.2	0.6	0.41	0.24	0.00	0.01	0.02
Coc2 - Coc2	1.51	2.32	-0.01	-1.04	-0.18	0.07	0.25	0.0	4.5	0.5	0.1	1.3	0.00	0.47	0.01	0.00	0.03
Coc3 - Coc3	0.92	4.46	-1.75	0.38	0.13	0.09	-0.05	4.5	0.4	0.2	0.1	0.0	0.69	0.03	0.00	0.00	0.00
							CONTRIBUTION CUMULEE =	6.1	6.5	0.8	0.4	1.9					
12 . Información Gremial																	
Gre1 - Gre1	2.78	0.80	0.61	0.35	0.06	-0.08	-0.08	1.7	0.9	0.1	0.2	0.2	0.47	0.15	0.00	0.01	0.01
Gre2 - Gre2	1.07	3.67	0.17	-1.25	-0.28	-0.01	0.24	0.0	4.6	0.9	0.0	0.9	0.01	0.43	0.02	0.00	0.02
Gre3 - Gre3	1.15	3.35	-1.63	0.32	0.12	0.19	-0.03	4.9	0.3	0.2	0.6	0.0	0.80	0.03	0.00	0.01	0.00
							CONTRIBUTION CUMULEE =	6.6	5.9	1.2	0.8	1.1					
13 . Notas Editoriales																	
Edi1 - Edi1	2.61	0.92	0.64	0.45	0.09	-0.11	-0.13	1.7	1.4	0.2	0.4	0.6	0.45	0.22	0.01	0.01	0.02
Edi2 - Edi2	1.32	2.79	0.09	-1.14	-0.29	0.06	0.33	0.0	4.7	1.2	0.1	2.0	0.00	0.46	0.03	0.00	0.04
Edi3 - Edi3	1.07	3.67	-1.67	0.31	0.15	0.18	-0.08	4.7	0.3	0.3	0.5	0.1	0.76	0.03	0.01	0.01	0.00
							CONTRIBUTION CUMULEE =	6.5	6.4	1.7	0.9	2.7					
14 . Comentarios económicos y políticos																	
Coe1 - Coe1	2.19	1.28	0.60	0.68	0.00	0.06	-0.15	1.2	2.8	0.0	0.1	0.7	0.28	0.36	0.00	0.00	0.02
Coe2 - Coe2	2.00	1.50	0.09	-0.90	0.00	-0.15	0.18	0.0	4.5	0.0	0.6	0.9	0.01	0.54	0.00	0.01	0.02
Coe3 - Coe3	0.81	5.19	-1.83	0.39	0.00	0.20	-0.03	4.3	0.3	0.0	0.4	0.0	0.65	0.03	0.00	0.01	0.00
							CONTRIBUTION CUMULEE =	5.6	7.6	0.0	1.1	1.6					
15 . Comentarios regionales																	
Cor1 - Cr1	1.60	2.13	0.65	0.83	0.00	-0.02	0.05	1.1	3.1	0.0	0.0	0.1	0.20	0.33	0.00	0.00	0.00
Cor2 - Cr2	2.52	0.99	0.21	-0.66	-0.03	-0.15	0.01	0.2	3.0	0.0	0.7	0.0	0.04	0.44	0.00	0.02	0.00
Cor3 - Cr3	0.89	4.65	-1.77	0.37	0.10	0.45	-0.13	4.4	0.3	0.1	2.3	0.2	0.68	0.03	0.00	0.04	0.00
							CONTRIBUTION CUMULEE =	5.7	6.4	0.1	3.0	0.3					
16 . Noticias Agropecuarias																	
Agr1 - Agr1	2.39	1.09	0.59	0.43	-0.14	-0.01	-0.14	1.3	1.3	0.5	0.0	0.7	0.32	0.17	0.02	0.00	0.02
Agr2 - Agr2	1.65	2.04	0.14	-0.86	0.23	-0.06	0.23	0.1	3.4	0.9	0.1	1.2	0.01	0.36	0.03	0.00	0.02
Agr3 - Agr3	0.96	4.19	-1.71	0.39	-0.05	0.13	-0.03	4.5	0.4	0.0	0.2	0.0	0.70	0.04	0.00	0.00	0.00
							CONTRIBUTION CUMULEE =	5.9	5.0	1.4	0.3	1.9					
17 . Noticias Regionales																	
Reg1 - Reg1	0.81	5.19	0.77	1.18	-0.07	0.14	0.52	0.8	3.1	0.0	0.2	3.0	0.12	0.27	0.00	0.00	0.05
Reg2 - Reg2	3.43	0.46	0.24	-0.37	0.02	0.01	-0.14	0.3	1.3	0.0	0.0	1.0	0.13	0.31	0.00	0.00	0.04
Reg3 - Reg3	0.76	5.57	-1.90	0.44	-0.02	-0.18	0.09	4.4	0.4	0.0	0.3	0.1	0.65	0.04	0.00	0.01	0.00
							CONTRIBUTION CUMULEE =	5.5	4.8	0.1	0.5	4.1					
18 . Investigación periodística																	
Inv1 - Inv1	1.96	1.56	0.66	0.75	0.12	-0.15	-0.12	1.4	3.0	0.3	0.6	0.4	0.28	0.36	0.01	0.01	0.01
Inv2 - Inv2	2.14	1.33	0.13	-0.85	-0.20	0.03	0.09	0.1	4.2	0.9	0.0	0.3	0.01	0.54	0.03	0.00	0.01
Inv3 - Inv3	0.90	4.55	-1.74	0.38	0.21	0.26	0.04	4.4	0.4	0.4	0.8	0.0	0.67	0.03	0.01	0.02	0.00
							CONTRIBUTION CUMULEE =	5.8	7.7	1.7	1.4	0.7					
19 . Información sobre Espectáculos																	
Esp1 - Esp1	2.25	1.22	0.60	0.56	0.30	-0.25	-0.03	1.3	2.0	2.2	1.8	0.0	0.30	0.26	0.07	0.05	0.00
Esp2 - Esp2	1.91	1.62	0.06	-0.84	-0.46	0.07	0.03	0.0	3.7	4.5	0.1	0.0	0.00	0.43	0.13	0.00	0.00
Esp3 - Esp3	0.84	4.96	-1.76	0.40	0.25	0.50	0.01	4.1	0.4	0.6	2.8	0.0	0.63	0.03	0.01	0.05	0.00
							CONTRIBUTION CUMULEE =	5.5	6.0	7.3	4.7	0.1					
20 . Sociales																	
Soc1 - Soc1	1.88	1.66	0.64	0.71	0.32	-0.33	-0.01	1.2	2.6	2.1	2.7	0.0	0.25	0.30	0.06	0.06	0.00
Soc2 - Soc2	2.19	1.28	0.20	-0.77	-0.38	0.13	0.01	0.1	3.6	3.5	0.5	0.0	0.03	0.46	0.11	0.01	0.00
Soc3 - Soc3	0.93	4.37	-1.77	0.37	0.24	0.36	0.00	4.6	0.4	0.6	1.6	0.0	0.71	0.03	0.01	0.03	0.00
							CONTRIBUTION CUMULEE =	6.0	6.5	6.2	4.7	0.0					
21 . Chistes, Horoscopos y entretenimiento																	
Chi1 - Chi1	1.91	1.62	0.50	0.63	0.40	-0.31	0.22	0.8	2.1	3.3	2.4	1.3	0.16	0.24	0.10	0.06	0.03
Chi2 - Chi2	2.58	0.94	0.03	-0.55	-0.36	0.18	-0.18	0.0	2.2	3.8	1.0	1.2	0.00	0.33	0.14	0.03	0.04
Chi3 - Chi3	0.51	8.76	-2.01	0.45	0.36	0.27	0.10	3.3	0.3	0.7	0.5	0.1	0.46	0.02	0.01	0.01	0.00
							CONTRIBUTION CUMULEE =	4.1	4.5	7.8	3.9	2.6					
22 . Suplemento y revista dominical																	
Sup1 - Sup1	1.82	1.75	0.59	0.76	0.31	-0.24	0.02	1.0	2.9	1.9	1.4	0.0	0.20	0.33	0.06	0.03	0.00
Sup2 - Sup2	2.38	1.10	0.16	-0.71	-0.34	0.01	-0.01	0.1	3.3	3.0	0.0	0.0	0.02	0.46	0.10	0.00	0.00
Sup3 - Sup3	0.81	5.19	-1.80	0.39	0.30	0.50	0.00	4.2	0.3	0.8	2.7	0.0	0.62	0.03	0.02	0.05	0.00
							CONTRIBUTION CUMULEE =	5.3	6.6	5.7	4.1	0.0					

Figura 6.7 Coordenadas, Contribuciones y cosenos cuadrados de las modalidades activas

Para el eje factorial 1 se puede separar Clasificados (contribución 5.7), Noticias de violencia, sexo y drogas (5.3), Información sobre corrupción política y social (5.0), Comentarios científicos (6.1), Información gremial (6.6), Notas editoriales (6.5), Noticias agropecuarias (5.9), Noticias regionales (5.5). Estas variables son las que más peso ejercen sobre el eje y son en las que se concentra la mayor cantidad de inercia.

Las modalidades de mayor peso dentro de cada variable son *lee* la sección ubicadas en el semieje positivo, y *no conoce* la sección

en el semieje negativo. Esto significa que en el primer eje factorial, que reúne el 31.38% de la inercia total de la nube de puntos, se oponen las personas que leen habitualmente diferentes secciones del diario con aquellas que ni siquiera la conocen.

En el factor 2 el análisis es similar; las variables relevantes son Noticias nacionales económicos (3.9), Noticias nacionales políticos (3.8), Comentarios económicos y políticos (7.6), Comentarios regionales (6.4), Investigación periodística (7.7), Información sobre espectáculos (6.0), Sociales (6.5) y Suplemento y revista dominical (6.6).

El segundo eje reúne el 18.86% de la inercia total de la nube de puntos; opone a las personas que leen determinadas secciones del diario de aquellas que no leen dichas secciones.

Puede observarse en Figura 12.8 que se agrupan las modalidades indicativas de individuos que leen secciones del diario y, cercana a ellas, los que no leen pero sí conocen que existen; más alejadas se encuentran aquellos individuos que no conocen las secciones del diario.

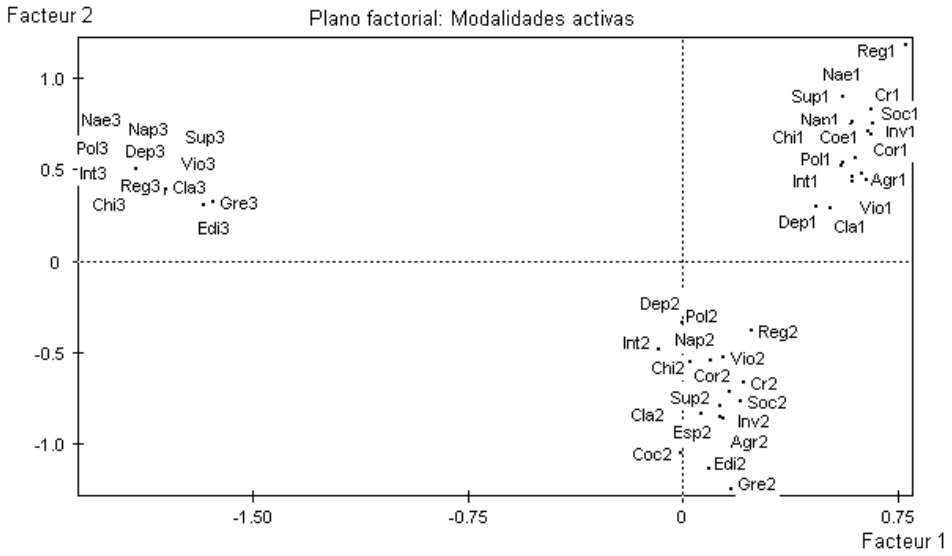


Figura 6.8

Algunas reagrupaciones son visibles sobre el primer plano, en particular el de comentarios científicos, información gremial, notas editoriales, comentarios económicos y políticos, comentarios regionales, noticias agropecuarias e investigación. La modalidad *leída* de estas secciones se sitúa en el primer cuadrante del plano, la modalidad *no leída* en el cuarto cuadrante y la modalidad *no conocida* mayormente en el segundo cuadrante.

El punto común de las diferentes secciones es su aspecto relativamente intelectual y de información general. Esto implica que interesan y no interesan a las mismas subpoblaciones. Se observa que la lectura de una de ellas se encuentra muy cercana a la lectura de otras; por ejemplo, comentarios económicos y políticos y comentarios regionales. El interés de un análisis multidimensional estriba en poner en evidencia tales fenómenos.

La calidad de representación de la modalidad *no lee* de la sección policiales es muy baja conjuntamente con la modalidad *lee* y *no*

lee de la sección deportes. Esto indica que su posición en el conjunto de las secciones está mal expresada sobre este plano. En realidad la variable policiales, con todas sus modalidades está bien representada en el quinto eje factorial; mientras que, deportes en otro eje más allá del quinto.

El factor 1 opone el conocimiento o no de las secciones del diario, mientras que el eje 2 opone la lectura de la sección a la no lectura. Si fuera necesario darle un nombre se podría decir que el eje 1 es "conocer" y el eje 2 es "leer".

Valores test o valores de prueba

Estos valores indican si la ubicación de las observaciones en el plano factorial es aleatoria o no, a través de probar la hipótesis nula de aleatoriedad en la construcción de cada eje factorial Rechazar la hipótesis nula significa que la proyección de la nube de puntos se ha realizado de manera no aleatoria. El valor test informado es el empírico y se contrasta con la distribución normal. Si el nivel de significatividad adoptado es de 0.05, el valor test debe superar a ∓ 1.96.

En la Figura 6.9 se encuentran los valores test o valores de prueba del estudio del diario regional

Se observa que, ha excepción de la modalidad 2, para la mayoría de las variables en el eje 1 todas las modalidades tienen valores test altos y en el eje 2 todas están por encima del nivel crítico; esto habla de una buena proyección de la nube de puntos sobre el plano factorial.

MODALITES	VALEURS-TEST	COORDONNEES	

IDEN - LIBELLE	EFF.	P.ABS	1	2	3	4	5	1	2	3	4	5	DISTO.
3 . Noticias Internacionales													
Int1 - Int1	128	128.00	8.0	7.6	-4.2	-1.1	-2.8	0.55	0.52	-0.29	-0.07	-0.20	1.52
Int2 - Int2	167	167.00	-1.6	-8.9	5.4	3.6	3.4	-0.09	-0.48	0.29	0.19	0.18	0.93
Int3 - Int3	27	27.00	-11.2	2.6	-2.4	-4.5	-1.0	-2.07	0.48	-0.45	-0.83	-0.19	10.93
4 . Clasificados													
Cla1 - Cla1	175	175.00	10.0	5.6	0.7	0.1	1.8	0.51	0.29	0.04	0.00	0.09	0.84
Cla2 - Cla2	91	91.00	1.4	-8.9	-0.1	-0.8	-2.7	0.13	-0.79	-0.01	-0.07	-0.24	2.54
Cla3 - Cla3	56	56.00	-14.8	3.2	-0.8	0.8	0.9	-1.81	0.39	-0.10	0.10	0.11	4.75
5 . Noticias de Violencia, sexo, droga													
Vio1 - Vio1	124	124.00	8.3	6.5	-3.3	2.4	11.0	0.59	0.46	-0.24	0.17	0.78	1.60
Vio2 - Vio2	146	146.00	2.3	-8.5	3.7	-2.6	-11.4	0.14	-0.52	0.23	-0.16	-0.70	1.21
Vio3 - Vio3	52	52.00	-14.2	2.9	-0.6	0.4	0.8	-1.80	0.36	-0.08	0.06	0.10	5.19
6 . Noticias policiales													
Pol1 - Pol1	91	91.00	6.2	6.0	-1.7	2.0	12.7	0.56	0.54	-0.15	0.18	1.13	2.54
Pol2 - Pol2	208	208.00	-0.3	-7.0	3.6	1.3	-13.4	-0.01	-0.29	0.15	0.05	-0.56	0.55
Pol3 - Pol3	23	23.00	-10.3	2.5	-3.7	-5.9	2.8	-2.08	0.50	-0.74	-1.18	0.57	13.00
7 . Noticias nacionales económicas													
Nae1 - Nae1	74	74.00	5.5	8.8	-9.1	6.4	-2.2	0.56	0.90	-0.93	0.66	-0.23	3.35
Nae2 - Nae2	228	228.00	0.0	-9.4	12.2	-0.2	1.9	0.00	-0.34	0.44	-0.01	0.07	0.41
Nae3 - Nae3	20	20.00	-9.5	2.3	-7.1	-10.7	0.4	-2.07	0.50	-1.53	-2.32	0.08	15.10
8 . Noticias nacionales políticas													
Nap1 - Nap1	90	90.00	6.5	8.3	-8.6	6.2	-4.3	0.59	0.75	-0.77	0.55	-0.39	2.58
Nap2 - Nap2	207	207.00	-0.5	-9.4	11.3	-1.3	4.3	-0.02	-0.39	0.47	-0.05	0.18	0.56
Nap3 - Nap3	25	25.00	-10.0	2.9	-5.8	-8.1	-0.5	-1.93	0.55	-1.12	-1.55	-0.10	11.88
9 . Información sobre corrupción política y social													
Cor1 - Cor1	104	104.00	8.1	8.6	-4.7	3.7	0.0	0.65	0.69	-0.38	0.30	0.00	2.10
Cor2 - Cor2	172	172.00	1.8	-10.5	5.4	-4.1	-0.2	0.09	-0.55	0.28	-0.21	-0.01	0.87
Cor3 - Cor3	46	46.00	-13.4	3.4	-1.5	0.8	0.3	-1.83	0.47	-0.20	0.11	0.05	6.00
10 . Deportes													
Dep1 - Dep1	122	122.00	6.5	4.2	-1.0	3.4	-2.3	0.47	0.30	-0.07	0.24	-0.16	1.64
Dep2 - Dep2	174	174.00	-0.8	-5.5	2.4	0.3	2.4	-0.04	-0.29	0.12	0.02	0.12	0.85
Dep3 - Dep3	26	26.00	-10.1	2.7	-2.5	-6.6	-0.3	-1.90	0.50	-0.47	-1.24	-0.07	11.38
34 . Nivel vivienda													
C1 - Media Superior	29	29.00	0.9	-2.9	1.3	0.3	3.6	0.17	-0.51	0.24	0.04	0.65	10.10
C2 - Media media	114	114.00	3.5	-2.2	0.6	-1.5	0.7	0.27	-0.17	0.05	-0.11	0.05	1.82
D1 - Media baja	118	118.00	-4.9	2.7	-0.3	0.1	-2.0	-0.36	0.20	-0.02	0.01	-0.15	1.73
D2 - Muy modesta/Marginal	61	61.00	1.0	1.4	-1.4	1.5	-1.0	0.11	0.16	-0.16	0.18	-0.12	4.28

Figura 6.9 Coordenadas y valores test de las modalidades

Los individuos

El análisis factorial permite superponer a la representación de las modalidades una representación de individuos.

En esta nueva nube de puntos hay que considerar dos aspectos: dos individuos están próximos si coinciden en las modalidades (o en el mayor número de ellas) y un individuo se situará en el centro de gravedad de las modalidades que lo caracterizan (o cercano a él). Coordenadas positivas en el primer plano para un individuo, lo sitúan asociado a las modalidades que caracterizan el primer cuadrante del plano de coordenadas factoriales. Este punto de vista, del análisis de los individuos, se confunde con el estudio de las modalidades suplementarias.

Ejemplo 6.5. En el estudio del diario regional, el AFCM superpone a la nube de modalidades, la nube de los 322 individuos encuestados. Dos individuos están próximos si leen las mismas secciones y un individuo se sitúa en el centro de gravedad de las modalidades lee, no lee o no conoce de las secciones que lee, no lee o no conoce.

Esto significa que, si un individuo tiene coordenadas positivas en el primer plano será asiduo lector de editoriales, artículos científicos, noticias gremiales, comentarios económicos y políticos e investigaciones.

Pero la posición de tal o cual punto no interesa en particular: el único interés del gráfico es ver que los individuos se reparten bastante uniformemente sobre el plano y que no hay clases de perfiles de lectura muy marcados. Por el contrario la posición de los encuestados interesa para representar las tendencias de quién en la pregunta ¿quién lee qué?.

Es aquí donde interviene la caracterización de los encuestados ya que se conoce para cada individuo su sexo, su nivel de estudios, su nivel socioeconómico, etc.

Para observar mejor como estas categorías están relacionadas con los modos de lectura, es posible representar los dos baricentros de varones y mujeres, los nueve baricentros de nivel de estudios, los trece de la actividad desarrollada, los cuatro de nivel socioeconómico de la vivienda, etc.

Las variables suplementarias: la caracterización

La proyección de las modalidades de estas variables sobre el plano indica que la variable relacionada con el primer factor es empleo y con el segundo educación.

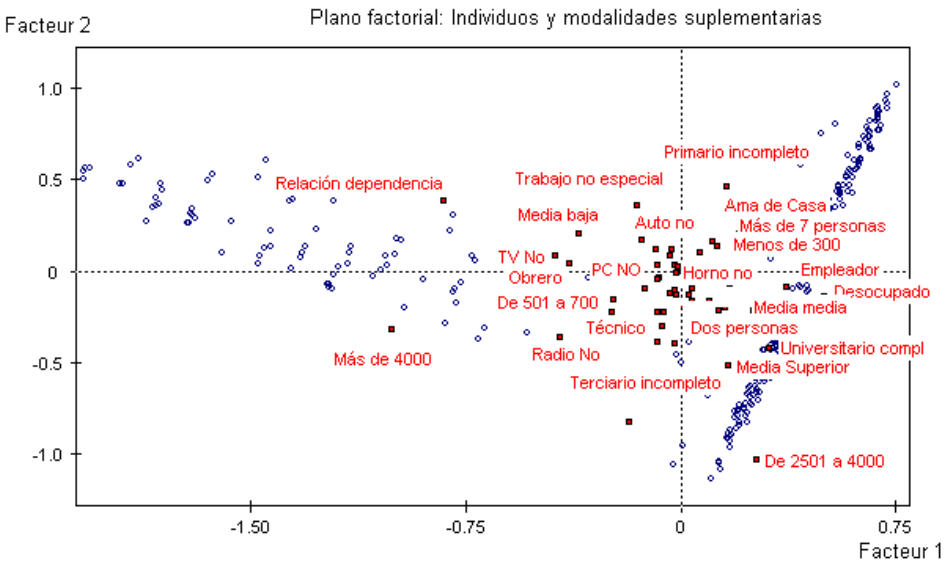


Figura 6.10

El nivel de estudios está muy ligado al perfil de lectura. Los lectores cuyo nivel de estudios es débil se oponen a los encuestados de nivel de estudios elevado. Estos últimos leen las informaciones “intelectuales” y pasan sin detenerse por las páginas de información gremial, de editoriales, y de otras informaciones más anecdóticas.

El respeto del orden de los niveles de estudios muestra que cuanto más instruido se es más “selectiva" se vuelve la lectura del diario regional: las personas de baja instrucción leen todo el diario, las de alta eligen qué leer. Bajo nivel de estudio se relaciona con noticias y comentarios regionales, suplementos y noticias nacionales económicas. A medida que se avanza en el nivel de estudios adquieren más relevancia las noticias políticas nacionales y las noticias internacionales.

Que los puntos representantes del nivel de estudios estén sobre una misma línea indica que, en la dispersión de la nube de los perfiles de lectura, existe otra dimensión independiente del nivel de estudios. Esta dimensión representada sobre el plano por la primera bisectriz de los dos ejes opone los individuos que leen un gran número de secciones a los que leen pocas.

c. Clasificación

El objetivo de hacer análisis factorial es encontrar tipologías de individuos y asociación de variables o características relacionadas con esos individuos. Luego de realizar AFCM, puede ser útil un procedimiento que se denomina clasificación y que consiste en reunir a los individuos de acuerdo a su similitud; es decir, formar grupos.

Se conocen dos grandes familias de métodos estadísticos que permiten clasificar un conjunto de observaciones:

- los métodos que fraccionan un conjunto de unidades de observación en subconjuntos homogéneos.
- los procedimientos que distribuyen o designan los elementos de un conjunto de observaciones en clases preestablecidas

Cualquiera de los métodos está compuesto de procedimientos destinados a definir clases de individuos similares entre sí. Las clases son subconjuntos de individuos que se caracterizan por la proximidad en el espacio, no significa que sean idénticos sino que comparten predominantemente ciertas características.

Cómo los métodos de análisis factorial, los métodos de clasificación jerárquica son destinados a producir una representación gráfica de la información contenida en la tabla de datos. Las clasificaciones jerárquicas tienen por objeto representar de manera sintética, el resultado de las comparaciones entre los objetos de una tabla de observaciones.

Clasificación jerárquica

Una clasificación jerárquica es una serie de particiones encajadas. Por cada nudo que se forma se calcula el índice de unión que permite evaluar la distancia entre los objetos clasificados. Entre los algoritmos de clasificación existentes se encuentran el método del vecino más

próximo, el método de los centroides de la distancia media y el método basado en el crecimiento mínimo del momento de orden dos en las clases de particiones indexadas.

El método del *vecino más próximo* consiste en evaluar las distancias entre todos los individuos. Si la distancia tiende a cero, estos individuos tienen similitud. De acuerdo a estas distancias, los individuos comienzan a agruparse formando los nudos y, estos nudos o grupos de individuos, se agrupan con otro nudo o grupos de individuos.

Cada nudo formado es sometido a una dócima a efectos de conocer su aleatoriedad. El tener valor test muy alto conduce, como en casos anteriores, a rechazar la hipótesis de aleatoriedad. Este proceso de agrupamiento de los individuos se representa gráficamente en un Dendograma. Esta gráfica permite observar sólo los grupos más representativos -es decir, las últimas 50 formaciones- e identificar la cantidad de clases en las que se divide la nube de puntos.

Ejemplo 6.6. En el análisis del diario regional, la Figura 6.11 muestra el resultado de los 50 nudos, o agrupaciones, más importantes de la clasificación jerárquica realizada sobre los 10 primeros ejes factoriales aplicando el método del vecino más próximo.

El nudo 594 se forma con el nudo 551 (primogénito) y el 478 (benjamín) quienes reúnen 8 individuos. El nudo 625 (18 individuos) está compuesto por el 603 (10 individuos) y el 604 (8 individuos). Luego el 625 y el 594 forman el 628 (26 individuos = 8+18). Sucesivamente se van agrupando, de acuerdo a la similitud existente entre ellos, hasta llegar al nudo 643 que reúne a los 322 individuos.

Cada nudo formado es sometido a un contraste a efectos de conocer su aleatoriedad. El tener valor test muy alto conduce, como en casos anteriores, a rechazar la hipótesis de aleatoriedad. Este análisis se realiza a los elementos ubicados en la base del dendograma y puede observarse en la Figura 6.12

En la Figura 6.14 se describe la formación de cada nudo del dendograma, el cual está ilustrado en la Figura 6.15. Se observa que la partición de la nube de puntos puede significar la constitución de 3 clases.

CLASSIFICATION HIERARCHIQUE (VOISINS RECIPROQUES)
SUR LES 10 PREMIERS AXES FACTORIELS
DESCRIPTION DES 50 NOEUDS D'INDICES LES PLUS ELEVES

NUM.	AINE	BENJ	EFF.	POIDS	INDICE	HISTOGRAMME DES INDICES DE NIVEAU
594	551	478	8	8.00	0.00233	*
595	488	527	12	12.00	0.00234	*
596	539	565	7	7.00	0.00234	*
597	517	555	5	5.00	0.00243	*
598	577	124	4	4.00	0.00244	*
599	522	534	7	7.00	0.00258	*
600	564	480	11	11.00	0.00260	*
601	588	540	11	11.00	0.00260	*
602	521	487	6	6.00	0.00268	*
603	448	574	10	10.00	0.00269	*
604	547	572	8	8.00	0.00289	*
605	570	519	20	20.00	0.00289	*
606	598	553	6	6.00	0.00306	*
607	589	578	18	18.00	0.00308	*
608	417	581	6	6.00	0.00321	*
609	562	582	14	14.00	0.00326	*
610	513	579	15	15.00	0.00336	*
611	558	597	8	8.00	0.00374	*
612	610	595	27	27.00	0.00407	*
613	593	590	9	9.00	0.00409	*
614	605	536	30	30.00	0.00425	*
615	608	557	9	9.00	0.00426	*
616	550	511	29	29.00	0.00491	*
617	587	544	11	11.00	0.00493	*
618	573	599	17	17.00	0.00533	*
619	585	583	31	31.00	0.00544	*
620	592	602	14	14.00	0.00582	*
621	596	580	9	9.00	0.00633	*
622	615	584	14	14.00	0.00687	*
623	614	591	46	46.00	0.00757	**
624	600	607	29	29.00	0.00768	**
625	603	604	18	18.00	0.00793	**
626	613	611	17	17.00	0.00813	**
627	609	626	31	31.00	0.00835	**
628	625	594	26	26.00	0.00840	**
629	576	606	11	11.00	0.00850	**
630	620	619	45	45.00	0.00995	**
631	586	618	33	33.00	0.01043	**
632	601	630	56	56.00	0.01310	**
633	621	617	20	20.00	0.01495	***
634	616	632	85	85.00	0.01561	***
635	627	629	42	42.00	0.02102	***
636	631	612	60	60.00	0.02243	****
637	628	624	55	55.00	0.02601	****
638	633	622	34	34.00	0.03474	*****
639	623	636	106	106.00	0.05335	********
640	638	635	76	76.00	0.06902	**********
641	634	637	140	140.00	0.07247	**********
642	639	641	246	246.00	0.25261	**********
643	640	642	322	322.00	0.56270	**********

SOMME DES INDICES DE NIVEAU = 1.47214

Figura 6.11 Clasificación de la nube de puntos

ELEMENTS				VALEURS-TEST					COORDONNEES				
NUM .	IDENT	POIDS	EFF	1	2	3	4	5	1	2	3	4	5
1 .	578	6.00	6	1.11	-1.09	1.38	-0.69	-1.48	0.45	-0.44	0.56	-0.28	-0.60
2 .	589	12.00	12	1.33	-2.07	1.99	-1.20	-2.68	0.38	-0.59	0.57	-0.34	-0.76
3 .	480	4.00	4	0.80	-1.45	-0.24	-0.10	1.82	0.40	-0.72	-0.12	-0.05	0.91
4 .	564	7.00	7	1.12	-1.90	-1.48	0.75	1.91	0.42	-0.71	-0.55	0.28	0.71
5 .	594	8.00	8	1.88	1.37	-0.63	-0.46	-5.45	0.66	0.48	-0.22	-0.16	-1.91
6 .	572	4.00	4	1.13	-0.36	-2.67	1.07	2.48	0.56	-0.18	-1.33	0.53	1.23
7 .	547	4.00	4	1.34	0.79	-2.56	1.22	-0.44	0.67	0.39	-1.27	0.61	-0.22
8 .	574	7.00	7	1.57	0.13	-3.30	1.93	-3.92	0.59	0.05	-1.24	0.72	-1.47
9 .	448	3.00	3	1.24	0.98	-3.96	2.43	-0.90	0.72	0.56	-2.28	1.40	-0.52
10 .	583	15.00	15	1.05	-4.98	-0.01	-0.17	-2.32	0.27	-1.26	0.00	-0.04	-0.59
11 .	585	16.00	16	0.93	-5.17	0.29	-0.63	-0.47	0.23	-1.26	0.07	-0.15	-0.11
12 .	487	2.00	2	0.61	-0.76	-1.33	0.64	-0.95	0.43	-0.53	-0.94	0.45	-0.67
13 .	521	4.00	4	0.86	-1.55	-3.54	1.35	-3.10	0.43	-0.77	-1.76	0.67	-1.54

14 . 592	8.00	8	0.93	-3.33	-2.03	1.02	1.47	0.32	-1.16	-0.71	0.36	0.51
15 . 540	4.00	4	0.71	-2.13	-1.32	0.42	4.07	0.36	-1.06	-0.66	0.21	2.03
16 . 588	7.00	7	0.82	-2.39	0.61	-0.36	5.30	0.31	-0.89	0.23	-0.13	1.99
17 . 511	4.00	4	0.61	-2.10	1.72	-1.41	-0.44	0.31	-1.04	0.86	-0.70	-0.22
18 . 550	25.00	25	0.75	-9.42	-1.34	-0.37	0.60	0.14	-1.81	-0.26	-0.07	0.12
19 . 527	5.00	5	1.27	0.49	3.77	-2.66	-1.81	0.56	0.22	1.68	-1.18	-0.81
20 . 488	7.00	7	1.58	1.28	4.45	-3.25	-2.68	0.59	0.48	1.67	-1.22	-1.00
21 . 579	6.00	6	1.59	1.74	2.75	-1.99	-2.65	0.64	0.70	1.11	-0.80	-1.07
22 . 513	9.00	9	2.02	2.06	4.08	-2.75	-3.22	0.66	0.68	1.34	-0.91	-1.06
23 . 534	4.00	4	0.95	-0.32	1.54	-1.07	3.22	0.47	-0.16	0.77	-0.53	1.60
24 . 522	3.00	3	1.03	0.30	1.53	-0.73	2.61	0.59	0.17	0.88	-0.42	1.50
25 . 573	10.00	10	2.16	1.94	1.78	-1.60	2.66	0.67	0.60	0.56	-0.50	0.83
26 . 586	16.00	16	3.12	4.43	3.54	-2.19	5.42	0.76	1.08	0.86	-0.54	1.32
27 . 591	16.00	16	3.38	5.08	-2.38	1.30	-4.03	0.83	1.24	-0.58	0.32	-0.98
28 . 536	10.00	10	2.75	3.96	-4.01	2.32	1.17	0.86	1.23	-1.25	0.72	0.37
29 . 519	9.00	9	2.70	4.71	-1.36	1.26	2.17	0.89	1.55	-0.45	0.41	0.71
30 . 570	11.00	11	2.90	4.83	-1.62	1.32	3.01	0.86	1.43	-0.48	0.39	0.89
31 . 553	2.00	2	-2.21	0.72	-0.05	2.54	-1.60	-1.56	0.51	-0.03	1.79	-1.13
32 . 200	1.00	1	-1.01	0.36	1.35	1.96	2.25	-1.01	0.36	1.35	1.96	2.25
33 . 577	3.00	3	-2.22	-0.21	2.76	0.56	-1.10	-1.28	-0.12	1.59	0.32	-0.63
34 . 576	5.00	5	-4.01	0.36	2.29	0.69	1.86	-1.78	0.16	1.02	0.31	0.82
35 . 555	2.00	2	-0.55	-0.59	0.50	0.57	0.73	-0.39	-0.42	0.35	0.40	0.52
36 . 517	3.00	3	-1.67	-0.69	1.59	2.15	-0.44	-0.96	-0.40	0.91	1.24	-0.26
37 . 558	3.00	3	-1.75	-0.48	0.83	0.13	-1.76	-1.01	-0.27	0.48	0.07	-1.01
38 . 590	5.00	5	-2.87	0.13	2.45	3.99	0.99	-1.27	0.06	1.09	1.77	0.44
39 . 593	4.00	4	-2.83	1.08	-1.99	4.60	0.14	-1.41	0.54	-0.99	2.29	0.07
40 . 582	6.00	6	-3.11	0.49	2.69	4.12	-1.56	-1.26	0.20	1.09	1.67	-0.63
41 . 562	8.00	8	-4.21	0.23	1.66	4.47	0.27	-1.47	0.08	0.58	1.56	0.10
42 . 584	5.00	5	-4.05	0.73	-3.64	-6.60	0.88	-1.80	0.32	-1.62	-2.93	0.39
43 . 557	3.00	3	-4.27	1.34	-3.09	-4.94	0.17	-2.46	0.77	-1.78	-2.84	0.10
44 . 581	4.00	4	-4.23	1.36	-4.36	-6.26	0.11	-2.10	0.68	-2.17	-3.11	0.05
45 . 417	2.00	2	-3.51	1.06	-1.34	-1.72	1.10	-2.48	0.75	-0.95	-1.21	0.78
46 . 544	4.00	4	-4.80	1.47	1.79	2.22	1.49	-2.39	0.73	0.89	1.10	0.74
47 . 587	7.00	7	-5.82	1.65	2.55	2.25	-1.10	-2.18	0.62	0.95	0.84	-0.41
48 . 580	2.00	2	-1.91	-0.18	-2.57	-1.76	-0.59	-1.35	-0.13	-1.82	-1.24	-0.41
49 . 565	4.00	4	-4.46	1.56	-1.31	-1.19	-0.60	-2.22	0.77	-0.65	-0.59	-0.30
50 . 539	3.00	3	-4.09	1.15	-1.18	-2.24	-0.81	-2.36	0.66	-0.68	-1.29	-0.46

Figura 6.13: Coordenadas y valores test de la base del dendograma

(INDICES EN POURCENTAGE DE LA SOMME DES INDICES : 1.33943)

NOEUD		SUCCESSEURS				COMPOSITION	
NUMERO	INDICE	AINE	BENJ	EFFECT.	POIDS	PREMIER	DERNIER
51	0.17	20	19	12	12.00	19	20
52	0.17	50	49	7	7.00	49	50
53	0.18	36	35	5	5.00	35	36
54	0.18	33	32	4	4.00	32	33
55	0.19	24	23	7	7.00	23	24
56	0.19	4	3	11	11.00	3	4
57	0.19	16	15	11	11.00	15	16
58	0.20	13	12	6	6.00	12	13
59	0.20	9	8	10	10.00	8	9
60	0.22	7	6	8	8.00	6	7
61	0.22	30	29	20	20.00	29	30
62	0.23	54	31	6	6.00	31	33
63	0.23	2	1	18	18.00	1	2
64	0.24	45	44	6	6.00	44	45
65	0.24	41	40	14	14.00	40	41
66	0.25	22	21	15	15.00	21	22
67	0.28	37	53	8	8.00	35	37
68	0.30	66	51	27	27.00	19	22
69	0.31	39	38	9	9.00	38	39
70	0.32	61	28	30	30.00	28	30
71	0.32	64	43	9	9.00	43	45
72	0.37	18	17	29	29.00	17	18
73	0.37	47	46	11	11.00	46	47
74	0.40	25	55	17	17.00	23	25
75	0.41	11	10	31	31.00	10	11
76	0.43	14	58	14	14.00	12	14
77	0.47	52	48	9	9.00	48	50
78	0.51	71	42	14	14.00	42	45
79	0.57	70	27	46	46.00	27	30
80	0.57	56	63	29	29.00	1	4
81	0.59	59	60	18	18.00	6	9
82	0.61	69	67	17	17.00	35	39
83	0.62	65	82	31	31.00	35	41
84	0.63	81	5	26	26.00	5	9
85	0.63	34	62	11	11.00	31	34
86	0.74	76	75	45	45.00	10	14
87	0.78	26	74	33	33.00	23	26
88	0.98	57	86	56	56.00	10	16
89	1.12	77	73	20	20.00	46	50
90	1.17	72	88	85	85.00	10	18
91	1.57	83	85	42	42.00	31	41
92	1.67	87	68	60	60.00	19	26
93	1.94	84	80	55	55.00	1	9

94	2.59	89	78	34	34.00	42	50
95	3.98	79	92	106	106.00	19	30
96	5.15	94	96	76	76.00	31	50
97	5.41	90	93	140	140.00	1	18
98	18.86	95	97	246	246.00	1	30
99	42.01	96	98	322	322.00	1	50

Figura 6.14: Descripción de los nodos de la jerarquía

Figura 6.15 Dendograma

d. Partición de la nube de puntos

Cuando se realiza la partición, en primer lugar se informa la cantidad de individuos que integran cada clase. Luego, esta partición es evaluada en términos de la inercia que involucra a cada clase. Al realizar la consolidación de la partición no necesariamente las clases quedan conformadas como en un comienzo. Esto depende de las distancias entre los centros de gravedad de cada clase y las distancias de los distintos individuos al centro de gravedad de la clase de pertenencia. Los individuos más cercanos al centro de gravedad de una clase son los individuos referentes de esa clase y se los considera característicos y representativos de todos los que la integran.

Estas clases, integradas por individuos, tienen asociadas ciertas características que quedan expuestas en la descripción de la partición.

Ejemplo 6.7. En el estudio del diario regional, la partición informa que

La clase 1 la integran 140 individuos que corresponden a los nudos 1 a 18, en la Figura 6.15 son los primeros 18 nudos identificados en la base.

La segunda clase reúne a 106 individuos ubicados desde la posición 19 a la 30.

La tercera clase corresponde a los 76 individuos reunidos desde la posición 31 a la 50.

El reagrupamiento asegura una mejor representación en el primer plano. Esto se ve en los valores test antes de la consolidación y los valores test posteriores a la consolidación.

Antes de la consolidación las clases tienen buena representación en el tercer eje factorial, luego de la consolidación las tres clases se representan bien en el primer plano factorial. La inercia total es la misma (1.4721), cambia la composición y la inercia existente dentro de cada grupo.

COUPURE 'a' DE L'ARBRE EN 3 CLASSES
FORMATION DES CLASSES (INDIVIDUS ACTIFS)
DESCRIPTION SOMMAIRE

CLASSE	EFFECTIF	POIDS	CONTENU
aa1a	140	140.00	1 A 18
aa2a	106	106.00	19 A 30
aa3a	76	76.00	31 A 50

COORDONNEES ET VALEURS-TEST AVANT CONSOLIDATION
AXES 1 A 5

	CLASSES		VALEURS-TEST					COORDONNEES					
IDEN - LIBELLE	EFF.	P.ABS	1	2	3	4	5	1	2	3	4	5	DISTO.
COUPURE 'a' DE L'ARBRE EN 3 CLASSES													
aa1a - CLASSE 1 / 3	140	140.00	5.5	-13.8	-4.4	0.8	-1.9	0.28	-0.53	-0.09	0.01	-0.03	0.37
aa2a - CLASSE 2 / 3	106	106.00	9.3	11.8	3.9	-2.8	2.1	0.59	0.57	0.09	-0.06	0.04	0.68
aa3a - CLASSE 3 / 3	76	76.00	-16.8	3.1	0.9	2.2	-0.1	-1.33	0.19	0.03	0.06	0.00	1.82

CONSOLIDATION DE LA PARTITION
AUTOUR DES 3 CENTRES DE CLASSES, REALISEE PAR 10 ITERATIONS A CENTRES MOBILES
PROGRESSION DE L'INERTIE INTER-CLASSES

```
+-----------+------------+------------+------------+
| ITERATION |   I.TOTALE |    I.INTER |  QUOTIENT  |
+-----------+------------+------------+------------+
|     0     |    1.47214 |    0.81531 |   0.55382  |
|     1     |    1.47214 |    0.85452 |   0.58046  |
|     2     |    1.47214 |    0.85684 |   0.58203  |
|     3     |    1.47214 |    0.85719 |   0.58228  |
|     4     |    1.47214 |    0.85726 |   0.58232  |
+-----------+------------+------------+------------+
ARRET APRES L'ITERATION  4 L'ACCROISSEMENT DE L'INERTIE INTER-CLASSES
PAR RAPPORT A L'ITERATION PRECEDENTE N'EST QUE DE 0.008 %.
DECOMPOSITION DE L'INERTIE
CALCULEE SUR  10 AXES.
+---------------+----------------+------------+------------------+-----------------+
|               |    INERTIES    |  EFFECTIFS |       POIDS      |    DISTANCES    |
| INERTIES      | AVANT    APRES | AVANT APRES |   AVANT    APRES |  AVANT   APRES  |
+---------------+----------------+------------+------------------+-----------------+
|               |                |            |                  |                 |
| INTER-CLASSES | 0.8153   0.8573 |            |                  |                 |
|               |                |            |                  |                 |
| INTRA-CLASSE  |                |            |                  |                 |
|               |                |            |                  |                 |
| CLASSE  1 /  3 | 0.2471   0.1747 |  140    124 |   140.00   124.00 | 0.3686   0.4711 |
| CLASSE  2 /  3 | 0.1528   0.1975 |  106    125 |   106.00   125.00 | 0.6840   0.6115 |
| CLASSE  3 /  3 | 0.2570   0.2426 |   76     73 |    76.00    73.00 | 1.8214   1.9339 |
|               |                |            |                  |                 |
|  TOTALE       | 1.4721   1.4721 |            |                  |                 |
+---------------+----------------+------------+------------------+-----------------+
QUOTIENT (INERTIE INTER / INERTIE TOTALE) : AVANT ... 0.5538
                                           APRES ... 0.5823
COORDONNEES ET VALEURS-TEST APRES CONSOLIDATION
AXES  1 A  5
+-----------------------------------------+---------------------------+---------------------------------+----------+
|                CLASSES                  |       VALEURS-TEST        |           COORDONNEES           |          |
|-----------------------------------------|---------------------------|---------------------------------|----------|
| IDEN - LIBELLE             EFF.   P.ABS |   1    2    3    4    5  |   1     2     3     4     5   | DISTO.   |
+-----------------------------------------+---------------------------+---------------------------------+----------+
|   COUPURE 'a' DE L'ARBRE EN  3 CLASSES                                                                |
|                                                                                                       |
| aa1a - CLASSE  1 /  3       124   124.00 |  4.0 -15.3 -1.7 -0.1  0.8 |  0.22 -0.65 -0.04  0.00  0.01 |    0.47 |
| aa2a - CLASSE  2 /  3       125   125.00 | 10.4  12.4  1.2 -1.6 -0.7 |  0.58  0.52  0.03 -0.03 -0.01 |    0.61 |
| aa3a - CLASSE  3 /  3        73    73.00 |-16.8   3.3  0.7  2.0 -0.1 | -1.37  0.20  0.02  0.06  0.00 |    1.93 |
+-----------------------------------------+---------------------------+---------------------------------+----------+
```

Figura 6.16 Partición del cuerpo del árbol jerárquico

En la Figura 6.17 se observan los individuos característicos de cada clase. Aquí solo se muestran los 10 principales que son los más cercanos al centro de gravedad.

En la Figura 6.18 se encuentra la primer clase. Los 124 individuos que la componen representan el 38.51% del total de individuos analizados. Se destacan por no leer las diferentes secciones del diario, son jefes de familia y pertenecen al nivel socioeconómico medio y medio superior.

```
PARANGONS
CLASSE   1/  3
EFFECTIF:  124
-------------------------------------------------------------------------------
|RG | DISTANCE  | IDENT. ||RG | DISTANCE  | IDENT. ||RG | DISTANCE  | IDENT. |
+---+-----------+--------++---+-----------+--------++---+-----------+--------+
|  1|    0.06893|256     ||  2|    0.06893|262     ||  3|    0.10484|99      |
|  4|    0.10484|257     ||  5|    0.10560|101     ||  6|    0.11096|295     |
|  7|    0.13495|30      ||  8|    0.15079|4       ||  9|    0.15780|384     |
| 10|    0.16321|302     ||   |           |        ||   |           |        |
+---+-----------+--------++---+-----------+--------++---+-----------+--------+
CLASSE   2/  3
EFFECTIF:  125
-------------------------------------------------------------------------------
|RG | DISTANCE  | IDENT. ||RG | DISTANCE  | IDENT. ||RG | DISTANCE  | IDENT. |
+---+-----------+--------++---+-----------+--------++---+-----------+--------+
```

RG	DISTANCE	IDENT.	RG	DISTANCE	IDENT.	RG	DISTANCE	IDENT.
1	0.21815	441	2	0.22005	436	3	0.22208	425
4	0.22396	69	5	0.23404	500	6	0.24538	493
7	0.24544	462	8	0.24969	214	9	0.25253	350
10	0.25335	371						

CLASSE 3/ 3
EFFECTIF: 73

RG	DISTANCE	IDENT.	RG	DISTANCE	IDENT.	RG	DISTANCE	IDENT.
1	0.38003	407	2	0.39686	66	3	0.40600	76
4	0.41637	248	5	0.53653	78	6	0.56249	207
7	0.60496	128	8	0.60496	340	9	0.62496	170
10	0.63522	219						

Figura 6.17. Individuos característicos de las clases

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				38.51	CLASSE 1 / 3		aa1a	124
13.12	0.000	78.26	87.10	42.86	2	Investigación periodística	Inv2	138
13.08	0.000	77.30	87.90	43.79	2	Sociales	Soc2	141
12.15	0.000	71.90	88.71	47.52	2	Suplemento y revista dominical	Sup2	153
11.98	0.000	77.52	80.65	40.06	2	Comentarios económicos y políticos	Coe2	129
11.04	0.000	67.28	87.90	50.31	2	Comentarios regionales	Cor2	162
10.95	0.000	94.20	52.42	21.43	2	Información Gremial	Gre2	69
10.84	0.000	75.61	75.00	38.20	2	Información sobre Espectáculos	Esp2	123
10.78	0.000	87.06	59.68	26.40	2	Notas Editoriales	Edi2	85
10.39	0.000	81.44	63.71	30.12	2	Comentarios Científicos	Coc2	97
10.07	0.000	55.20	98.39	68.63	2	Noticias Regionales	Reg2	221
9.50	0.000	75.47	64.52	32.92	2	Noticias Agropecuarias	Agr2	106
9.38	0.000	62.65	83.87	51.55	2	Chistes, Horoscopos y entretenimiento	Chi2	166
9.36	0.000	61.63	85.48	53.42	2	Información sobre corrupción política y social	Cor2	172
7.47	0.000	53.14	88.71	64.29	2	Noticias nacionales políticas	Nap2	207
7.27	0.000	60.27	70.97	45.34	2	Noticias de Violencia, sexo, droga	Vio2	146
7.18	0.000	50.44	92.74	70.81	2	Noticias nacionales económicas	Nae2	228
7.05	0.000	56.89	76.61	51.86	2	Noticias Internacionales	Int2	167
5.93	0.000	64.84	47.58	28.26	2	Clasificados	Cla2	91
5.26	0.000	49.04	82.26	64.60	2	Noticias policiales	Pol2	208
3.58	0.000	47.70	66.94	54.04	2	Deportes	Dep2	174
2.89	0.002	65.52	15.32	9.01	Media Superior	Nivel vivienda	C1	29
2.52	0.006	46.05	56.45	47.20	1	Es jefe de familia?	JefS	152
2.29	0.011	47.37	43.55	35.40	Media media	Nivel vivienda	C2	114
1.91	0.028	53.85	16.94	12.11	Jubilado/pensionado	Actividad desarrollada	Jub	39
1.75	0.040	52.50	16.94	12.42	Dos personas	Número de persoans en el hogar	Dos	40
1.64	0.050	56.00	11.29	7.76	Terciario completo	Nivel de educación	Teco	25
1.13	0.129	47.62	16.13	13.04	PC Si	Computadora	PCS	42
1.09	0.139	60.00	4.84	3.11	Una persona	Número de persoans en el hogar	Una	10
1.05	0.148	46.67	16.94	13.98	De 701 a 1000	Nivel de ingresos	850	45
1.00	0.158	52.94	7.26	5.28	Universitario incomp	Nivel de educación	Unin	17
1.00	0.159	41.72	50.81	46.89	Auto Si	Auto	AutS	151
0.93	0.175	46.34	15.32	12.73	Horno si	Horno a microondas	HMS	41
0.87	0.191	53.85	5.65	4.04	Terciario incompleto	Nivel de educación	Tein	13
0.77	0.222	42.86	26.61	23.91	Secundario completo	Nivel de educación	Seco	77
0.70	0.241	46.43	10.48	8.70	Estudiante	Actividad desarrollada	Est	28
0.70	0.241	46.43	10.48	8.70	De 1001 a 1500	Nivel de ingresos	1250	28
0.61	0.270	40.69	47.58	45.03	De 3 a 4 personas	Número de persoans en el hogar	3a4	145
0.55	0.292	41.18	33.87	31.68	Video Si	Video	VidS	102
0.45	0.328	50.00	4.03	3.11	Empleador	Actividad desarrollada	Empr	10
0.38	0.350	44.00	8.87	7.76	Profesional y otros	Actividad desarrollada	Prof	25
0.30	0.380	46.15	4.84	4.04	Técnico	Actividad desarrollada	Tec	13
0.03	0.490	38.89	50.81	50.31	1	Sexo	Sex1	162

Figura 6.18. Descripción de la partición. Clase 1.

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				38.82	CLASSE 2 / 3		aa2a	125
14.67	0.000	86.51	87.20	39.13	1	Investigación periodística	Inv1	126
14.18	0.000	86.78	84.00	37.58	1	Sociales	Soc1	121
14.09	0.000	72.62	97.60	52.17	1	Notas Editoriales	Edi1	168
13.94	0.000	72.89	96.80	51.55	1	Comentarios Científicos	Coc1	166
13.76	0.000	79.43	89.60	43.79	1	Comentarios económicos y políticos	Coe1	141
13.48	0.000	68.72	98.40	55.59	1	Información Gremial	Gre1	179
13.06	0.000	76.55	88.80	45.03	1	Información sobre Espectáculos	Esp1	145
12.77	0.000	83.76	78.40	36.34	1	Suplemento y revista dominical	Sup1	117
11.38	0.000	83.50	68.80	31.99	1	Comentarios regionales	Cor1	103
11.10	0.000	69.48	85.60	47.83	1	Noticias Agropecuarias	Agr1	154
10.99	0.000	76.42	75.20	38.20	1	Chistes, Horoscopos y entretenimiento	Chi1	123
10.70	0.000	80.77	67.20	32.30	1	Información sobre corrupción política y social	Cor1	104
10.18	0.000	72.66	74.40	39.75	1	Noticias Internacionales	Int1	128
9.49	0.000	96.15	40.00	16.15	1	Noticias Regionales	Reg1	52
8.84	0.000	77.78	56.00	27.95	1	Noticias nacionales políticas	Nap1	90
8.62	0.000	68.55	68.00	38.51	1	Noticias de Violencia, sexo, droga	Vio1	124

8.43	0.000	59.43	83.20	54.35	1	Clasificados	Cla1	175
8.11	0.000	79.73	47.20	22.98	1	Noticias nacionales económicas	Nae1	74
6.88	0.000	69.23	50.40	28.26	1	Noticias policiales	Pol1	91
5.45	0.000	58.20	56.80	37.89	1	Deportes	Dep1	122
2.15	0.016	52.73	23.20	17.08	Ama de Casa	Actividad desarrollada	Ama	55
2.01	0.022	60.87	11.20	7.14	Primario incompleto	Nivel de educación	Prin	23
1.77	0.038	47.19	33.60	27.64	Primario completo	Nivel de educación	Prco	89
1.73	0.042	51.02	20.00	15.22	Menos de 300	Nivel de ingresos	300	49
1.69	0.046	49.18	24.00	18.94	Muy modesta/Marginal	Nivel vivienda	D2	61
1.41	0.080	45.26	34.40	29.50	De 300 a 500	Nivel de ingresos	400	95
1.28	0.101	48.84	16.80	13.35	Más de 7 personas	Número de persoans en el hogar	+7	43
1.03	0.151	41.76	56.80	52.80	2	Es jefe de familia?	JefN	170
1.02	0.153	62.50	4.00	2.48	Desocupado	Actividad desarrollada	Des	8
1.02	0.155	42.98	39.20	35.40	Media media	Nivel vivienda	C2	114
0.75	0.225	42.86	28.80	26.09	De 5 a 6 personas	Número de persoans en el hogar	5a6	84
0.72	0.237	42.16	34.40	31.68	Video Si	Video	VidS	102
0.56	0.289	47.37	7.20	5.90	De 1501 a 2500	Nivel de ingresos	2000	19
0.48	0.314	40.35	55.20	53.11	Auto no	Auto	AutN	171
0.48	0.316	39.50	88.80	87.27	Horno no	Horno a microondas	HMN	281
0.35	0.363	44.00	8.80	7.76	Profesional y otros	Actividad desarrollada	Prof	25
0.26	0.396	39.29	88.00	86.96	PC NO	Computadora	PCN	280
0.14	0.443	40.30	21.60	20.81	Secundario incom	Nivel de educación	Sein	67
0.14	0.444	39.51	51.20	50.31	1	Sexo	Sex1	162
0.14	0.445	39.05	98.40	97.83	Radio Sí	Radio	RS	315
0.06	0.478	42.86	4.80	4.35	Trabajo no especial	Actividad desarrollada	Noes	14

Figura 6.19 Descripción de la partición. Clase 2.

El valor test docima la aleatoriedad de esa modalidad en la clase. Si el valor test es superior a 2 con un nivel de confianza del 95.44%, significa que se rechaza la hipótesis de aleatoriedad con lo cual esa modalidad es realmente una característica de la clase.

La modalidad *no lee* de la variable Investigación periodística tiene 138 puntos en total; es decir, el 42.86% del total de personas entrevistadas presentan esta modalidad.

Luego, de los 138 que presentan la modalidad en el total, el 78.26% (108 individuos) forman parte de esta clase lo que significa que el 87.10% de los individuos de esta clase *no lee* la sección Investigación periodística.

En la Figura 6.19 se encuentra la segunda clase que reúne el 38.82% de los individuos; los individuos de esta clase se caracterizan por leer las diferentes secciones del diario, ser ama de casa y tener un bajo nivel de instrucción. Estas características deben interpretarse como predominantes y no exclusivas; es decir, la mayoría de los que *leen* investigación periodística están en esta clase (el 87.20%), lo cual significa que puede haber en este grupo alguna persona que no lea esta sección o que no la conozca.

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				22.67	CLASSE 3 / 3		aa3a	73
16.13	0.000	93.24	94.52	22.98	3	Información Gremial	Gre3	74
15.38	0.000	94.20	89.04	21.43	3	Notas Editoriales	Edi3	69
14.77	0.000	98.33	80.82	18.63	3	Sociales	Soc3	60
14.34	0.000	95.16	80.82	19.25	3	Noticias Agropecuarias	Agr3	62
14.20	0.000	98.25	76.71	17.70	3	Comentarios regionales	Cor3	57
14.15	0.000	96.61	78.08	18.32	3	Comentarios Científicos	Coc3	59
13.74	0.000	100.00	71.23	16.15	3	Suplemento y revista dominical	Sup3	52
13.58	0.000	96.43	73.97	17.39	3	Clasificados	Cla3	56
13.56	0.000	94.83	75.34	18.01	3	Investigación periodística	Inv3	58
13.26	0.000	98.08	69.86	16.15	3	Noticias de Violencia, sexo, droga	Vio3	52
13.21	0.000	96.30	71.23	16.77	3	Información sobre Espectáculos	Esp3	54
12.84	0.000	96.15	68.49	16.15	3	Comentarios económicos y políticos	Coe3	52
12.71	0.000	97.96	65.75	15.22	3	Noticias Regionales	Reg3	49
11.72	0.000	95.65	60.27	14.29	3	Información sobre corrupción política y social	Cor3	46
10.26	0.000	100.00	45.21	10.25	3	Chistes, Horoscopos y entretenimiento	Chi3	33
9.09	0.000	100.00	36.99	8.39	3	Noticias Internacionales	Int3	27
8.33	0.000	96.15	34.25	8.07	3	Deportes	Dep3	26
8.27	0.000	100.00	31.51	7.14	3	Noticias policiales	Pol3	23
8.12	0.000	96.00	32.88	7.76	3	Noticias nacionales políticas	Nap3	25
7.00	0.000	95.00	26.03	6.21	3	Noticias nacionales económicas	Nae3	20
4.83	0.000	38.14	61.64	36.65	Media baja	Nivel vivienda	D1	118
1.95	0.026	32.35	30.14	21.12	De 501 a 700	Nivel de ingresos	600	68
1.85	0.032	50.00	8.22	3.73	Relación dependencia	Actividad desarrollada	Dep	12
1.77	0.039	37.93	15.07	9.01	Comerciante s/pers	Actividad desarrollada	Com	29
1.63	0.052	25.45	76.71	68.32	Video No	Video	VidN	220
1.59	0.056	26.47	61.64	52.80	2	Es jefe de familia?	JefN	170
1.34	0.090	28.57	32.88	26.09	De 5 a 6 personas	Número de persoans en el hogar	5a6	84
1.08	0.139	28.36	26.03	20.81	Secundario incom	Nivel de educación	Sein	67
0.96	0.169	40.00	5.48	3.11	Ns/Nc	Nivel de ingresos	Nsnc	10
0.88	0.189	35.71	6.85	4.35	Trabajo no especial	Actividad desarrollada	Noes	14
0.79	0.215	23.57	90.41	86.96	PC NO	Computadora	PCN	280
0.70	0.241	24.83	49.32	45.03	De 3 a 4 personas	Número de persoans en el hogar	3a4	145
0.59	0.278	33.33	5.48	3.73	Obrero	Actividad desarrollada	Obr	12
0.46	0.322	23.98	56.16	53.11	Auto no	Auto	AutN	171
0.44	0.331	23.95	54.79	51.86	2	Noticias Internacionales	Int2	167
0.33	0.372	23.75	52.05	49.69	2	Sexo	Sex2	160
0.29	0.385	23.13	89.04	87.27	Horno no	Horno a microondas	HMN	281
0.12	0.453	23.88	21.92	20.81	Empleado s/jerarquía	Actividad desarrollada	Empl	67
0.11	0.457	25.00	9.59	8.70	Estudiante	Actividad desarrollada	Est	28
0.01	0.497	28.57	2.74	2.17	Radio No	Radio	RN	7

Figura 6.20. Descripción de la partición. Clase 3.

La tercera clase se describe en la Figura 12.20. El 22.67% del total de individuos están en esta clase y se caracterizan por no conocer las diferentes secciones del diario, pertenecer a un nivel socioeconómico bajo y con ingresos promedios del hogar de $600.

Información gremial y notas editoriales son las secciones del diario que menos conoce la población estudiada.

e. Tabla de Burt

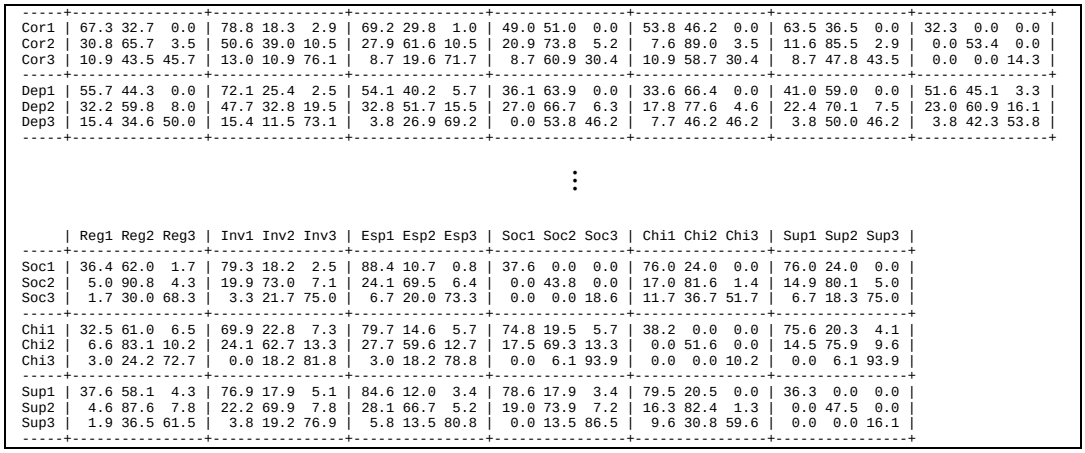
Esta tabla está formada por una yuxtaposición de tablas de contingencia definidas por las variables de todos dos grupos.

	Int1	Int2	Int3	Cla1	Cla2	Cla3	Vio1	Vio2	Vio3	Pol1	Pol2	Pol3	Nae1	Nae2	Nae3	Nap1	Nap2	Nap3	Cor1	Cor2	Cor3
Int1	128	0	0																		
Int2	0	167	0																		
Int3	0	0	27																		
Cla1	102	73	0	175	0	0															
Cla2	21	69	1	0	91	0															
Cla3	5	25	26	0	0	56															
Vio1	75	49	0	98	22	4	124	0	0												
Vio2	49	88	9	74	62	10	0	146	0												
Vio3	4	30	18	3	7	42	0	0	52												
Pol1	52	38	1	74	12	5	84	5	2	91	0	0									
Pol2	72	119	17	100	77	31	39	139	30	0	208	0									
Pol3	4	10	9	1	2	20	1	2	20	0	0	23									
Nae1	57	16	1	58	10	6	52	20	2	37	36	1	74	0	0						
Nae2	67	147	14	116	78	34	72	122	34	54	165	9	0	228	0						
Nae3	4	4	12	1	3	16	0	4	16	0	7	13	0	0	20						
Nap1	65	25	0	70	16	4	55	31	4	37	52	1	70	20	0	90	0	0			
Nap2	59	135	13	104	70	33	67	109	31	52	143	12	2	201	4	0	207	0			
Nap3	4	7	14	1	5	19	2	6	17	2	13	10	2	7	16	0	0	25			
Cor1	70	34	0	82	19	3	72	31	1	51	53	0	56	48	0	66	38	0	104	0	0
Cor2	53	113	6	87	67	18	48	106	18	36	127	9	13	153	6	20	147	5	0	172	0
Cor3	5	20	21	6	5	35	4	9	33	4	28	14	5	27	14	4	22	20	0	0	46
Dep1	68	54	0	88	31	3	66	49	7	44	78	0	41	81	0	50	72	0	63	55	4
Dep2	56	104	14	83	57	34	57	90	27	47	116	11	31	135	8	39	122	13	40	106	28
Dep3	4	9	13	4	3	19	1	7	18	0	14	12	2	12	12	1	13	12	1	11	14
Coc1	107	59	0	132	32	2	97	67	2	70	96	0	61	105	0	74	91	1	88	76	2
Coc2	16	77	4	35	49	13	23	63	11	17	75	5	8	86	3	12	80	5	12	74	11
Coc3	5	31	23	8	10	41	4	16	39	4	37	18	5	37	17	4	36	19	4	22	33
Gre1	105	74	0	139	38	2	103	76	0	73	106	0	61	118	0	75	103	1	90	88	1
Gre2	16	52	1	25	41	3	16	51	2	12	56	1	6	61	2	9	59	1	9	56	4
Gre3	7	41	26	11	12	51	5	19	50	6	46	22	7	49	18	6	45	23	5	28	41
Edi1	101	67	0	132	35	1	96	72	0	70	98	0	62	106	0	75	92	1	88	79	1
Edi2	21	61	3	33	44	8	26	55	4	16	67	2	6	76	3	10	72	3	14	65	6
Edi3	6	39	24	10	12	47	2	19	48	5	43	21	6	46	17	5	43	21	2	28	39
⋮	⋮																				
Esp1	47	94	4	110	32	3	145	0	0												
Esp2	4	107	12	15	97	11	0	123	0												
Esp3	1	20	33	1	9	44	0	0	54												
Soc1	44	75	2	96	22	3	107	13	1	121	0	0									
Soc2	7	128	6	28	103	10	34	98	9	0	141	0									
Soc3	1	18	41	2	13	45	4	12	44	0	0	60									
Chi1	40	75	8	86	28	9	98	18	7	92	24	7	123	0	0						
Chi2	11	138	17	40	104	22	46	99	21	29	115	22	0	166	0						
Chi3	1	8	24	0	6	27	1	6	26	0	2	31	0	0	33						
Sup1	44	68	5	90	21	6	99	14	4	92	21	4	93	24	0	117	0	0			
Sup2	7	134	12	34	107	12	43	102	8	29	113	11	25	126	2	0	153	0			
Sup3	1	19	32	2	10	40	3	7	42	0	7	45	5	16	31	0	0	52			

Figura 6.21 Tabla de Burt – Perfiles absolutos

LE TRI-A-PLAT DE CHAQUE QUESTION FIGURE SUR LA DIAGONALE CORRESPONDANTE
TOUS LES NOMBRES SONT EXPRIMES EN POURCENTAGES

	Int1	Int2	Int3	Cla1	Cla2	Cla3	Vio1	Vio2	Vio3	Pol1	Pol2	Pol3	Nae1	Nae2	Nae3	Nap1	Nap2	Nap3	Cor1	Cor2	Cor3
Int1	39.8	0.0	0.0	79.7	16.4	3.9	58.6	38.3	3.1	40.6	56.2	3.1	44.5	52.3	3.1	50.8	46.1	3.1	54.7	41.4	3.9
Int2	0.0	51.9	0.0	43.7	41.3	15.0	29.3	52.7	18.0	22.8	71.3	6.0	9.6	88.0	2.4	15.0	80.8	4.2	20.4	67.7	12.0
Int3	0.0	0.0	8.4	0.0	3.7	96.3	0.0	33.3	66.7	3.7	63.0	33.3	3.7	51.9	44.4	0.0	48.1	51.9	0.0	22.2	77.8
Cla1	58.3	41.7	0.0	54.3	0.0	0.0	56.0	42.3	1.7	42.3	57.1	0.6	33.1	66.3	0.6	40.0	59.4	0.6	46.9	49.7	3.4
Cla2	23.1	75.8	1.1	0.0	28.3	0.0	24.2	68.1	7.7	13.2	84.6	2.2	11.0	85.7	3.3	17.6	76.9	5.5	20.9	73.6	5.5
Cla3	8.9	44.6	46.4	0.0	0.0	17.4	7.1	17.9	75.0	8.9	55.4	35.7	10.7	60.7	28.6	7.1	58.9	33.9	5.4	32.1	62.5
Vio1	60.5	39.5	0.0	79.0	17.7	3.2	38.5	0.0	0.0	67.7	31.5	0.8	41.9	58.1	0.0	44.4	54.0	1.6	58.1	38.7	3.2
Vio2	33.6	60.3	6.2	50.7	42.5	6.8	0.0	45.3	0.0	3.4	95.2	1.4	13.7	83.6	2.7	21.2	74.7	4.1	21.2	72.6	6.2
Vio3	7.7	57.7	34.6	5.8	13.5	80.8	0.0	0.0	16.1	3.8	57.7	38.5	3.8	65.4	30.8	7.7	59.6	32.7	1.9	34.6	63.5
Pol1	57.1	41.8	1.1	81.3	13.2	5.5	92.3	5.5	2.2	28.3	0.0	0.0	40.7	59.3	0.0	40.7	57.1	2.2	56.0	39.6	4.4
Pol2	34.6	57.2	8.2	48.1	37.0	14.9	18.8	66.8	14.4	0.0	64.6	0.0	17.3	79.3	3.4	25.0	68.8	6.2	25.5	61.1	13.5
Pol3	17.4	43.5	39.1	4.3	8.7	87.0	4.3	8.7	87.0	0.0	0.0	7.1	4.3	39.1	56.5	4.3	52.2	43.5	0.0	39.1	60.9
Nae1	77.0	21.6	1.4	78.4	13.5	8.1	70.3	27.0	2.7	50.0	48.6	1.4	23.0	0.0	0.0	94.6	2.7	2.7	75.7	17.6	6.8
Nae2	29.4	64.5	6.1	50.9	34.2	14.9	31.6	53.5	14.9	23.7	72.4	3.9	0.0	70.8	0.0	8.8	88.2	3.1	21.1	67.1	11.8
Nae3	20.0	20.0	60.0	5.0	15.0	80.0	0.0	20.0	80.0	0.0	35.0	65.0	0.0	0.0	6.2	0.0	20.0	80.0	0.0	30.0	70.0
Nap1	72.2	27.8	0.0	77.8	17.8	4.4	61.1	34.4	4.4	41.1	57.8	1.1	77.8	22.2	0.0	28.0	0.0	0.0	73.3	22.2	4.4
Nap2	28.5	65.2	6.3	50.2	33.8	15.9	32.4	52.7	15.0	25.1	69.1	5.8	1.0	97.1	1.9	0.0	64.3	0.0	18.4	71.0	10.6
Nap3	16.0	28.0	56.0	4.0	20.0	76.0	8.0	24.0	68.0	8.0	52.0	40.0	8.0	28.0	64.0	0.0	0.0	7.8	0.0	20.0	80.0



	Cor1	Cor2	Cor3	Dep1	Dep2	Dep3
Cor1	67.3 32.7 0.0	78.8 18.3 2.9	69.2 29.8 1.0	49.0 51.0 0.0	53.8 46.2 0.0	63.5 36.5 0.0	32.3 0.0 0.0
Cor2	30.8 65.7 3.5	50.6 39.0 10.5	27.9 61.6 10.5	20.9 73.8 5.2	7.6 89.0 3.5	11.6 85.5 2.9	0.0 53.4 0.0
Cor3	10.9 43.5 45.7	13.0 10.9 76.1	8.7 19.6 71.7	8.7 60.9 30.4	10.9 58.7 30.4	8.7 47.8 43.5	0.0 0.0 14.3
Dep1	55.7 44.3 0.0	72.1 25.4 2.5	54.1 40.2 5.7	36.1 63.9 0.0	33.6 66.4 0.0	41.0 59.0 0.0	51.6 45.1 3.3
Dep2	32.2 59.8 8.0	47.7 32.8 19.5	32.8 51.7 15.5	27.0 66.7 6.3	17.8 77.6 4.6	22.4 70.1 7.5	23.0 60.9 16.1
Dep3	15.4 34.6 50.0	15.4 11.5 73.1	3.8 26.9 69.2	0.0 53.8 46.2	7.7 46.2 46.2	3.8 50.0 46.2	3.8 42.3 53.8

⋮

	Reg1	Reg2	Reg3	Inv1	Inv2	Inv3	Esp1	Esp2	Esp3	Soc1	Soc2	Soc3	Chi1	Chi2	Chi3	Sup1	Sup2	Sup3
Soc1	36.4	62.0	1.7	79.3	18.2	2.5	88.4	10.7	0.8	37.6	0.0	0.0	76.0	24.0	0.0	76.0	24.0	0.0
Soc2	5.0	90.8	4.3	19.9	73.0	7.1	24.1	69.5	6.4	0.0	43.8	0.0	17.0	81.6	1.4	14.9	80.1	5.0
Soc3	1.7	30.0	68.3	3.3	21.7	75.0	6.7	20.0	73.3	0.0	0.0	18.6	11.7	36.7	51.7	6.7	18.3	75.0
Chi1	32.5	61.0	6.5	69.9	22.8	7.3	79.7	14.6	5.7	74.8	19.5	5.7	38.2	0.0	0.0	75.6	20.3	4.1
Chi2	6.6	83.1	10.2	24.1	62.7	13.3	27.7	59.6	12.7	17.5	69.3	13.3	0.0	51.6	0.0	14.5	75.9	9.6
Chi3	3.0	24.2	72.7	0.0	18.2	81.8	3.0	18.2	78.8	0.0	6.1	93.9	0.0	0.0	10.2	0.0	6.1	93.9
Sup1	37.6	58.1	4.3	76.9	17.9	5.1	84.6	12.0	3.4	78.6	17.9	3.4	79.5	20.5	0.0	36.3	0.0	0.0
Sup2	4.6	87.6	7.8	22.2	69.9	7.8	28.1	66.7	5.2	19.0	73.9	7.2	16.3	82.4	1.3	0.0	47.5	0.0
Sup3	1.9	36.5	61.5	3.8	19.2	76.9	5.8	13.5	80.8	0.0	13.5	86.5	9.6	30.8	59.6	0.0	0.0	16.1

Figura 6.22. Tabla de Burt - Perfiles horizontales

La Tabla de Burt consiste en la confección de la totalidad de tablas de contingencia posibles. Se construye de dos maneras: de puntos absolutos y de perfiles horizontales. En la tabla de valores absolutos (Figura 6.21) definiendo una modalidad (leen la sección noticias internacionales, Int1) podemos leer hacia abajo que 102 personas leen los clasificados (Cla1). Es decir, 102 personas leen conjuntamente noticias internacionales y clasificadas.

La Tabla de perfiles horizontales indica cómo se distribuye una modalidad en términos de otra variable. Por ejemplo: de los que leen noticias internacionales, el 79.7% lee los clasificados, el 16.4% no los lee y el 3.9% no conoce la sección. Cuidado!!!, la lectura sólo es válida horizontalmente, la lectura vertical carece de sentido.

6.2. ANÁLISIS DE COMPONENTES PRINCIPALES

El análisis de componentes principales se utiliza cuando las variables objeto de estudio son cuantitativas. En líneas generales, el análisis es similar al de factorial de correspondencias múltiples. Particularmente, las diferencias se encuentran en el centro de gravedad de la clase –ubicado en el punto donde confluyen las medias de todas las variables-, la matriz de inercia –matriz de correlaciones de todas las variables-, la dimensión de la nube de puntos variables –la cantidad de variables activas-.

El Análisis de Componentes Principales se realiza con el software SPAD. En lo que sigue se presenta la manera de transformar el formato Excel en formato SPAD y se instruye en la generación de estadísticas descriptivas, el análisis de componentes principales, la clasificación de la base y la partición y descripción de la nube de puntos.

Se dispone de información proveniente del Censo Nacional de Población y Vivienda 2001 de Argentina; el archivo Población2001.xls puede obtenerse de http://www.econometricos.com.ar/cursos-de-posgrado/ieunrc/basesdatos/.

En el archivo Poblacion2001.xlxs, la solapa de *Valores relativos* tiene la tabla de datos cuyo detalle se observa en la Figura 6.23.

Identificador	Descripción de la variable
Provincia	
Departamento	Primera Subdivisión de las Provincias
Población	Para el año 2001 en cantidad de habitantes
Crecimiento	Variación relativa 2001/1991 en porcentaje
Superficie	En km cuadrados
Densidad	Número de habitantes por Km cuadrado
PH	Cantidad de personas que vive en hogares respecto del total de población en el departamento, en porcentaje

Figura 12.23 Continúa…

Figura 12.23 Continuación

PIC	Cantidad de personas que vive en instituciones colectivas respecto del total de población en el departamento, en porcentaje
P10mas	Cantidad de personas de 10 años o más respecto del total de población del Departamento, en porcentajes
Alfabetos	Cantidad de personas alfabetizadas respecto del total de personas de 10 años o más por departamento, en porcentajes
Analfabetos	Cantidad de personas analfabetas respecto del total de personas de 10 años o más por departamento, en porcentajes
NBIH	Hogares con Necesidades Básicas Insatisfechas, en porcentaje
NBIP	Personas con Necesidades Básicas Insatisfechas, en porcentaje
Hogares	Cantidad de Hogares que viven en viviendas
TVcasa	Cantidad de hogares que viven en Casas respecto de los de hogares que viven en viviendas, en porcentaje
TVrancho	Cantidad de hogares que viven en Ranchos respecto de los de hogares que viven en viviendas, en porcentaje

TVcasilla	Cantidad de hogares que viven en Casillas respecto de los de hogares que viven en viviendas, en porcentaje
Tvdepartamento	Cantidad de hogares que viven en Departamentos respecto de los de hogares que viven en viviendas, en porcentaje
Tvinq	Cantidad de hogares que viven en Piezas de inquilinato respecto de los de hogares que viven en viviendas, en porcentaje
Tvhotel	Cantidad de hogares que viven en Piezas de hotel o pensión respecto de los de hogares que viven en viviendas, en porcentaje
Tvlocal	Cantidad de hogares que viven en Local no construido para habitación respecto de los de hogares que viven en viviendas, en porcentaje
Tvmovil	Cantidad de hogares que viven en Vivienda móvil respecto de los de hogares que viven en viviendas, en porcentaje
Población	Cantidad de Población que vive en viviendas
Pcasa	Cantidad de Población que vive en Casas respecto de la cantidad de población que vive en viviendas, en porcentaje
Prancho	Cantidad de Población que vive en Rancho respecto de la cantidad de población que vive en viviendas, en porcentaje
Pcasilla	Cantidad de Población que vive en Casilla respecto de la cantidad de población que vive en viviendas, en porcentaje
Pdepartamento	Cantidad de Población que vive en Departamento respecto de la cantidad de población que vive en viviendas, en porcentaje
Pinq	Cantidad de Población que vive en Piezas de Inquilinato respecto de la cantidad de población que vive en viviendas, en porcentaje
Photel	Cantidad de Población que vive en Piezas en hotel o pensión respecto de la cantidad de población que vive en viviendas, en porcentaje
Plocal	Cantidad de Población que vive en Local no construido para habitación respecto de la cantidad de población que vive en viviendas, en porcentaje
Pmovil	Cantidad de Población que vive en Vivienda móvil respecto de la cantidad de población que vive en viviendas, en porcentaje
HIDA1	Cantidad de hogares que cuentan con Inodoro con descarga de agua y desagüe a red pública respecto del total de hogares que viven en viviendas, en porcentaje

Figura 6.23 Continúa…

Figura 6.23 Continuación

HIDA2	Cantidad de hogares que cuentan con Inodoro con descarga de agua y a desagüe a cámara séptica y pozo ciego respecto del total de hogares que viven en viviendas, en porcentaje
HIDA3	Cantidad de hogares que cuentan con Inodoro con descarga de agua y desagüe a pozo ciego u hoyo, excavación en la tierra, etc. respecto del total de hogares que viven en viviendas, en porcentaje
HISDA	Cantidad de hogares que cuentan con Inodoro sin descarga de agua o sin inodoro respecto del total de hogares que viven en viviendas, en porcentaje
CALMATI_HIDA1	Cantidad de hogares con calidad de materiales tipo I y cuentan con Inodoro con descarga de agua y desagüe a red pública respecto del total de hogares que viven en viviendas, en porcentaje
CLAMATI_HIDA2	Cantidad de hogares con calidad de materiales tipo I y cuentan con Inodoro con descarga de agua y a desagüe a cámara séptica y pozo ciego respecto del total de hogares que viven en viviendas, en porcentaje

CALMATI_HIDA3	Cantidad de hogares con calidad de materiales tipo I y cuentan con Inodoro con descarga de agua y desagüe a pozo ciego u hoyo, excavación en la tierra, etc. respecto del total de hogares que viven en viviendas, en porcentaje
CALMATI_HISDA	Cantidad de hogares con calidad de materiales tipo I y cuentan con Inodoro sin descarga de agua o sin inodoro respecto del total de hogares que viven en viviendas, en porcentaje
PISO1	Cantidad de hogares con pisos de Cerámica, baldosa, mosaico, mármol, madera o alfombrado respecto del total de hogares que viven en viviendas, en porcentaje
PISO2	Cantidad de hogares con pisos de Cemento o ladrillo fijo respecto del total de hogares que viven en viviendas, en porcentaje
PISO3	Cantidad de hogares con pisos de Tierra o ladrillo suelto respecto del total de hogares que viven en viviendas, en porcentaje
PISO4	Cantidad de hogares con pisos de otro tipo respecto del total de hogares que viven en viviendas, en porcentaje

Figura 6.23 Variables integrantes de la base de datos

En la solapa *Valores absolutos* se encuentran las mismas variables pero en cantidades observadas en cada jurisdicción; en la solapa *Identificadores* el detalle precedente de las variables y en la solapa *Referencias* comentarios aportados por el Instituto Nacional de Estadísticas y Censos (INDEC) a la base de datos del Censo Nacional de Población y Viviendas de 2001.

De lenguaje Excel a lenguaje SPAD

La tabla disponible en formato Excel debe transformarse en formato SPAD, previo paso por formato texto. Es oportuno mencionar que el procedimiento a seguir, en cada uno de los software, se realizó en el marco de Windows XP, el cambio a otro sistema operativo puede presentar modificaciones.

Desde Excel 2010, para guardar la tabla de datos en formato texto el camino es *Archivo-Guardar como-Guardar como tipo-Texto (delimitado por tabulaciones) (*.txt)* (Figura 1). Aparecerá un mensaje que advierte sobre la pérdida de las hojas no activas; dicho de otro modo, sólo se guarda la hoja de cálculo a la vista porque el formato txt no reconoce múltiples hojas; este mensaje debe aceptarse (Figura 2). Nuevamente aparece un mensaje acerca de la no compatibilidad entre el archivo generado y el formato en el cual se generó; a la pregunta de

Excel se debe responder SI (Figura 3). Al querer cerrar el programa, pregunta si se quiere volver a guardarlo, se selecciona *Guardar* (Figura 4). Nuevamente, aparece el mensaje sobre la no compatibilidad, al igual que antes se selecciona SI (Figura 5).

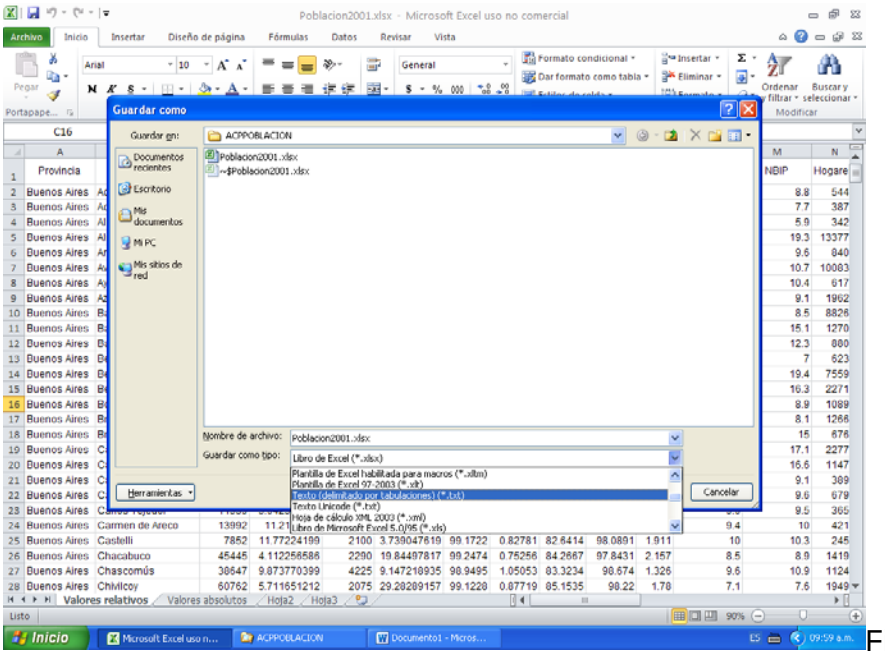


Figura 1

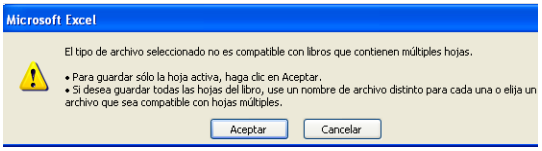


Figura 2

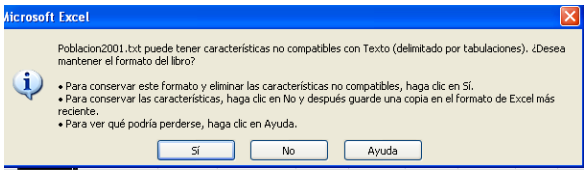


Figura 3

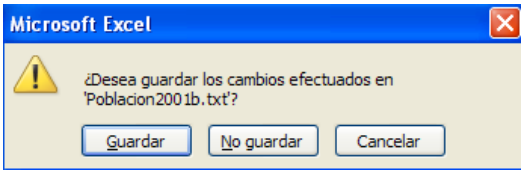


Figura 4

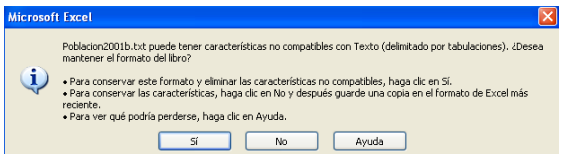


Figura 5

Desde *Inicio-Todos los programas-Accesorios-WordPad* (Figura 6), se ingresa al editor de texto Word Pad con el cual puede verse - desde *Archivo-Abrir* y localizando el archivo de texto en el directorio (Figura 7)- el archivo generado (Figura 8).

Desde *Inicio-Todos los Programas-Decisia-Spad v56en* (Figura 9) se presenta la pantalla de ingreso a SPAD (Figura 10). En el extremo inferior derecho hay una flecha que permite ingresar al entorno del software, el que cuenta con 3 bloques: la barra horizontal con acceso *Dataset/Chain/Tools/Options/Window/Help*; la ventana *Methods* y la ventana *Chain* (Figura 11). En la primera son de uso frecuente *Dataset y Chain;* la segunda es útil cuando se ha identificado el logo de una herramienta, cuando es así no es necesario seguir el procedimiento formal sino se le indica a través del logo lo que se requiere; la última ventana es donde se desarrollan los análisis en SPAD. Se observa un diagrama de flujo con ícono gris seguido del ícono con la palabra END.

Importar archivo

Para importar un archivo de texto desde SPAD se procede desde *Dataset/Import/Import Ascii file* (Figura 12), se despliega una ventana en la que se selecciona *New* para que habilite el cuadro de diálogo *IMPORT NAME*, aquí se ingresa el nombre del archivo a importar sin extensión: *Poblacion2001* (Figura 13). Al aceptar, se despliega el explorador de Windows que permite localizar el archivo *Poblacion2001.txt* (Figura 14) y abrirlo.

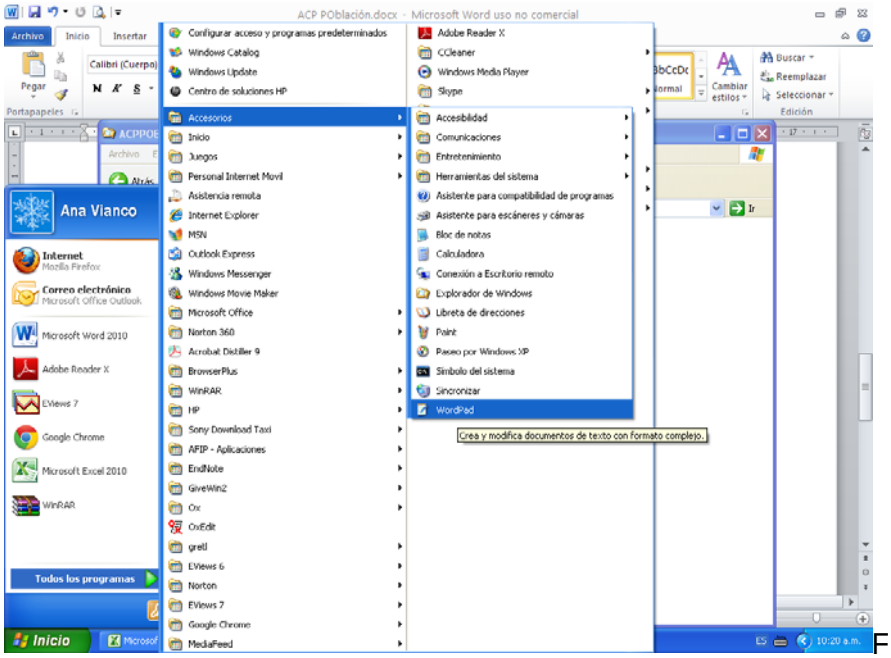


Figura 6

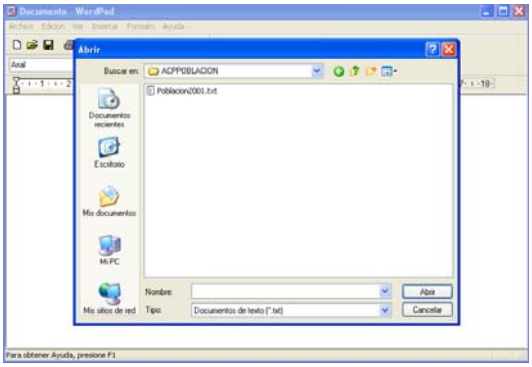


Figura 7

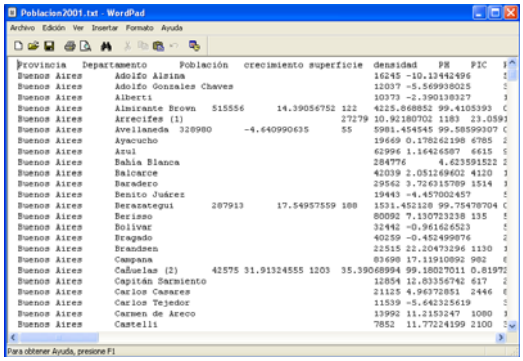


Figura 8

La acción anterior da paso a la ventana de importación de SPAD en la cual hay que verificar que se encuentren de manera correcta las indicaciones (Figura 15). En primer lugar, si el formato es fijo (*Fixed*) o delimitado por tabulaciones (*Delimited*); si lo números están expresados en formato decimal con punto (*dot*) o coma (*comma*); si la tabla de datos tiene la primer fila con rótulos para las columnas debe estar seleccionada la leyenda *The first line has the variable label*; la última opción *Several lines for a case* es necesaria cuando la cantidad de columnas hace que, por cada observación, se tenga más de una línea. En el cuadro inferior de la ventana se observa una muestra de la base de datos a importar. Presionando *Next* se accede al paso siguiente.

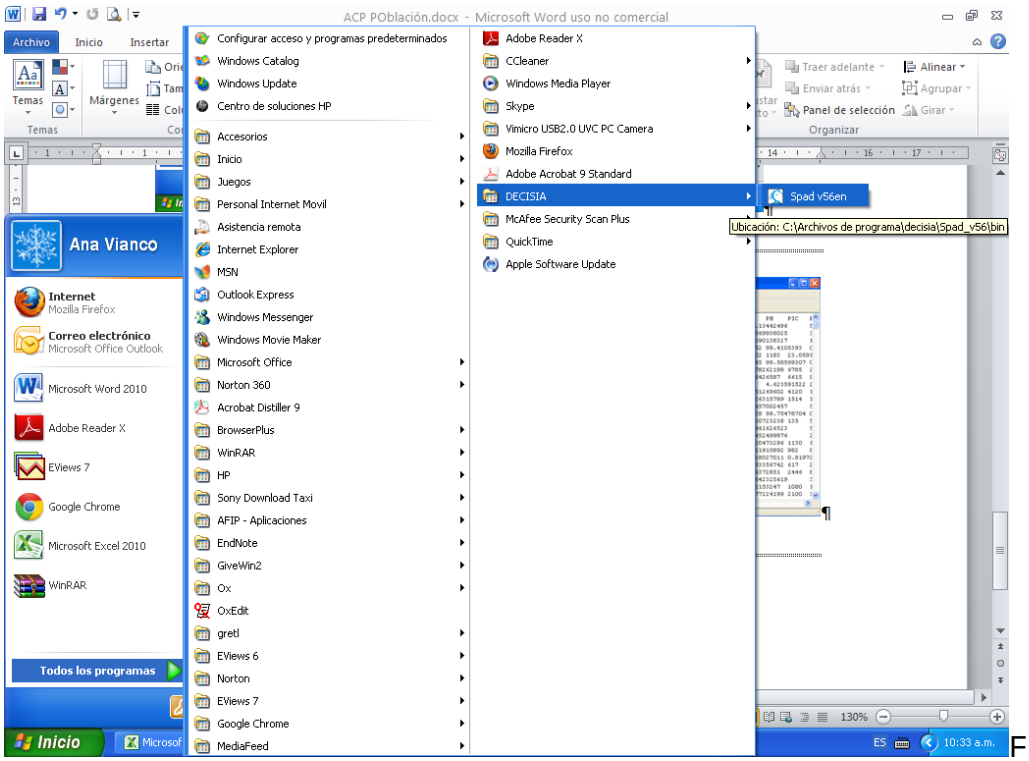


Figura 9

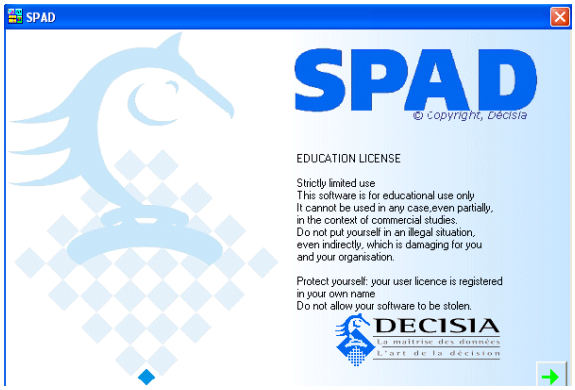


Figura 10

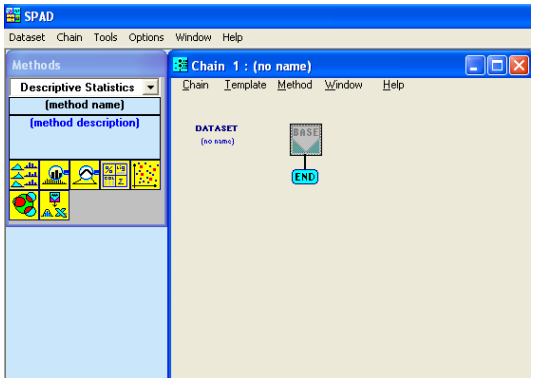


Figura 11

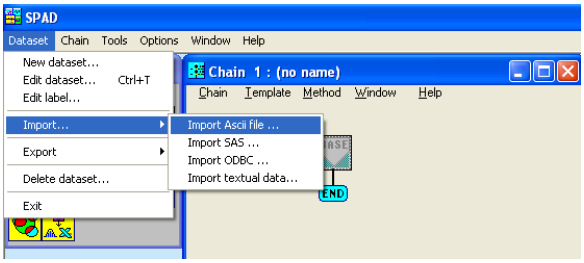


Figura 12

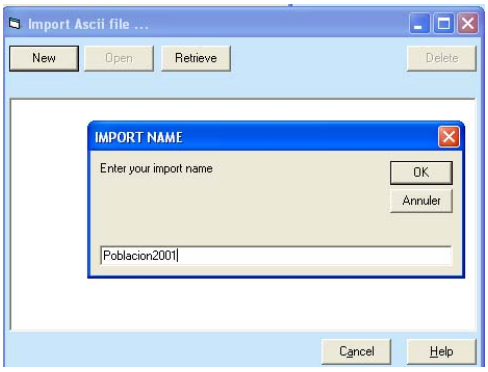


Figura 13

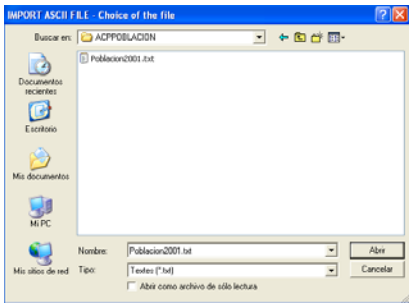


Figura 14

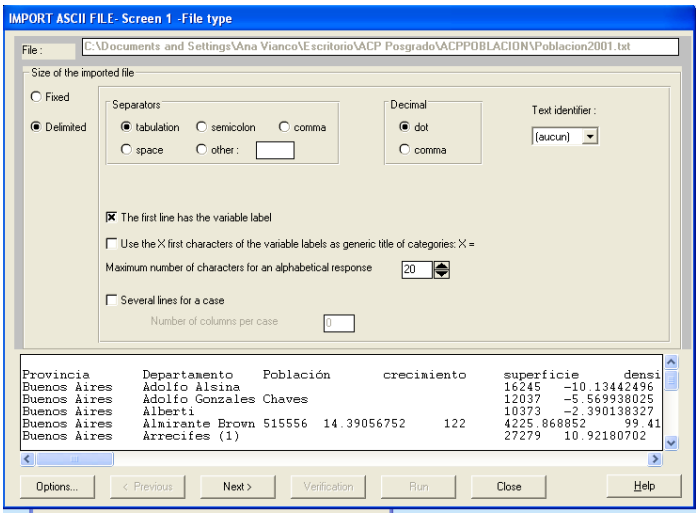


Figura 15

La segunda etapa en el proceso de importación tiene una matriz donde, en cada fila hay una variable y en las columnas se tiene la posición de la variable en la base de datos (*Column*), el tipo de variable (*Type*) y el identificador de la variable (*Variable label*) (Figura 16). El tipo de variable debe asignarse a cada una de ellas en esta etapa; para la variable Provincia que es cualitativa y tiene los nombres de cada una de ellas, el tipo de variable a seleccionar es *Alphabetical* por intermedio del cual SPAD generará una categoría por cada jurisdicción. El Departamento es la unidad de observación por esto se le asigna el tipo *Identifier*. Las variables de análisis son cuantitativas continuas, por esto el tipo es *Continuos*. Presionando *Run* se despliega la ventana del explorador de Windows para localizar el directorio donde guardar la base generada. La extensión del archivo base de SPAD es *.sba* (Figura 17); cuando se le indica que guarde, procesa la información y genera el reporte de control indicando las variables existentes en la base (Figura 18).

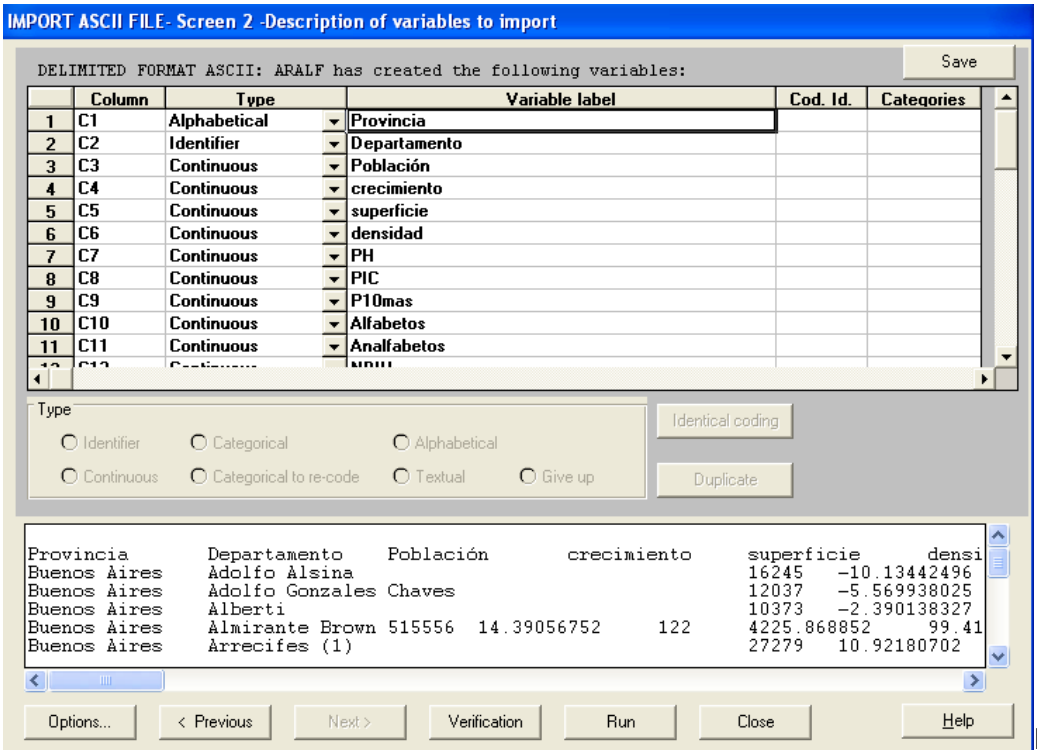


Figura 16

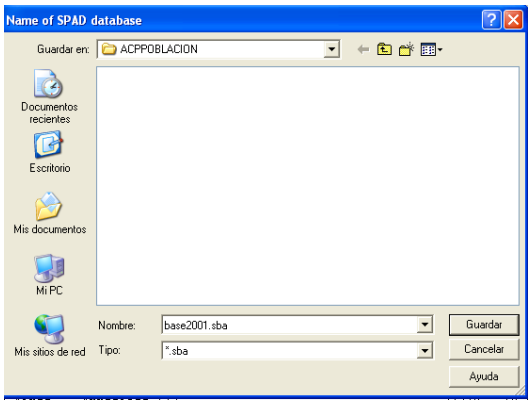


Figura 17

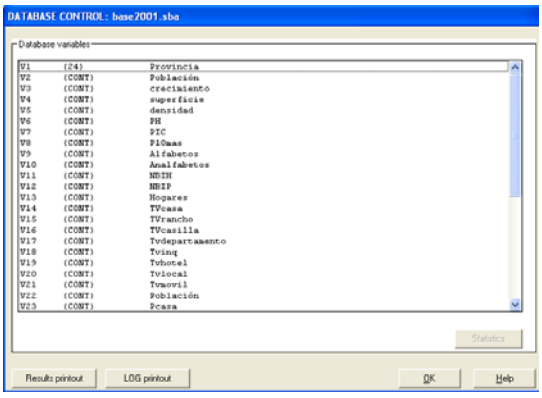


Figura 18

Control de la base importada

En el menú principal de SPAD, desde *Dataset=Edit dataset…* (Figura 19) se selecciona la base generada en el paso anterior (Figura 20). En este entorno, SPAD permite incorporar una buena descripción de las

variables y modificar el nombre asignado por defecto a las categorías de las variables cualitativas; también es posible realizar modificaciones en los datos sin necesidad de volver a transformar la base (Figura 21). Aquí hay tres ventanas: la de Variables, la de Valores y la de Categorías. En la de Variables se observa el tipo de variable y su recorrido expresado en los valores mínimo y máximo. En la de Valores se reproduce la tabla de datos. En la de categorías están los identificadores de cada modalidad de la variable cualitativa.

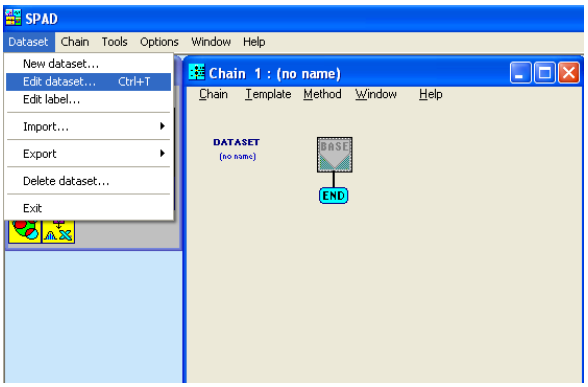


Figura 19

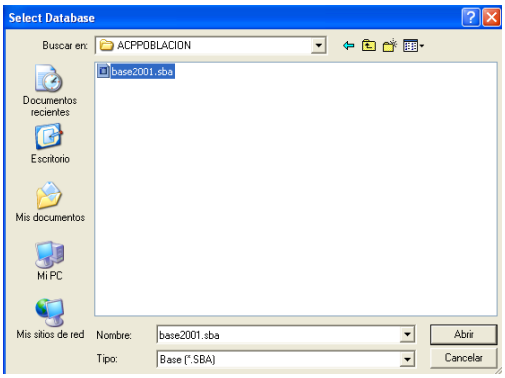


Figura 20

Trabajar con la base

En primer lugar es necesario que la base se encuentre en un directorio generado sólo para el trabajo desde SPAD. Este software tiene la característica de generar muchos archivos vinculados entre los que hay temporales, de texto, de Excel y otros formatos propios de SPAD; esto hace que no sea posible identificar su pertenencia, si la base está sola en una carpeta, todo lo que se genere allí tiene vinculación con ella.

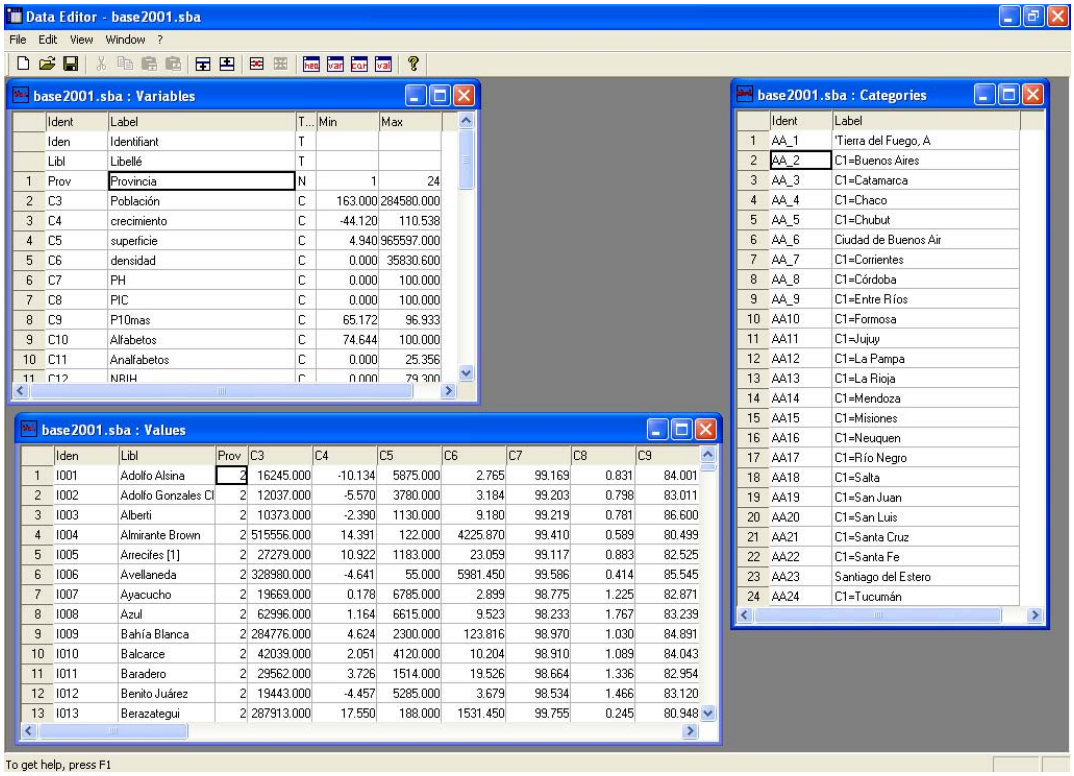


Figura 21

Para comenzar a trabajar, se sitúa en la ventana *Chain 1: (no name)* y se procede desde *Chain-Selected dataset…* seleccionando el archivo (Figura 22). Luego, para realizar un análisis, deben seguirse los siguientes pasos:

1) Desde Method-Insert method (Figura 23).

2) Desde Method-Selected method (Figura 24), para realizar una estadística descriptiva de las variables contenidas en la base, se selecciona Descriptive Statistics en la primera ventana y Marginal distributions, histograms en la segunda ventana (Figura 25); al aceptar esta etapa, se observa que en la ventana Chain el ícono gris ahora tiene ilustraciones (Figura 26).

3) Desde Method-Parameters (Figura 27) se le indica sobre cuáles variables se requiere el análisis; por defecto, la ventana se abre en la

solapa Marginal distributions indicando en el primer cuadro (Available variables) las variables cualitativas disponibles; si se quiere conocer la frecuencia con la que aparecen cada una de las modalidades de las variables cualitativas, debe seleccionarse la variable y enviarse al cuadro siguiente (Selected variables) con el uso de los íconos de flechas (Figura 28). La solapa Histograms Caracterisation permite seleccionar las variables cuantitativas para el análisis descriptivo (Figura 29). Las siguientes solapas están diseñadas por defecto para un análisis general de toda la base, particularmente en Casos pueden seleccionar sólo algunos individuos para que formen parte del análisis y en Weighting se puede asignar ponderaciones a las observaciones. Al aceptar las indicaciones (con OK) el ícono adquiere colores, esto significa que se han definido los parámetros para que el método se ejecute (Figura 30).

4) Desde Method-Run method… se le indica a SPAD que debe ejecutar la solicitud de análisis realizada en los pasos anteriores; aparecerá un mensaje en el que pregunta si se quiere guardar el entorno de trabajo (Chain) (Figura 31), se debe responder SI. Entonces, se despliega el explorador de Windows para localizar un directorio donde guardarlo, no es conveniente aceptar la indicación brindada por el software, se recomienda guardarlo en el directorio donde se encuentra la tabla de datos en Excel y la tabla de datos en formato texto (Figura 32). Al indicarle Guardar, se despliega un cuadro de diálogo donde brinda la posibilidad de asignar una leyenda al entorno de trabajo, la cual no es obligatoria (Figura 33). Al dar OK se completa el paso de generación del análisis, a continuación del ícono sobre el cual se estaba trabajando, aparecen un ícono texto, un ícono gráfico y un ícono planilla (Figura 34).

El ícono texto contiene el resultado del análisis (Figura 35) del cual puede verse la versión completa en el Anexo Estadísticas Descriptivas.

El procedimiento de ir a *Method-Insert method*, *Method Select*, *Method Parameters* y *Method Run* se repite para todos los análisis realizados en SPAD.

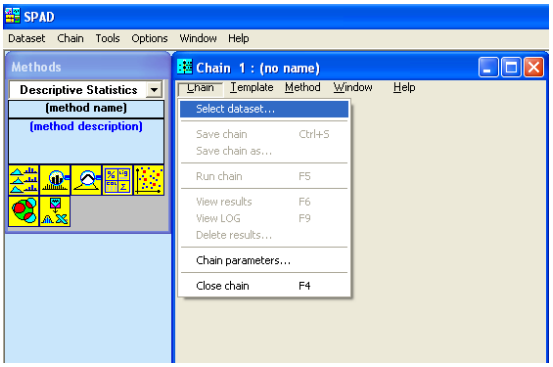


Figura 22

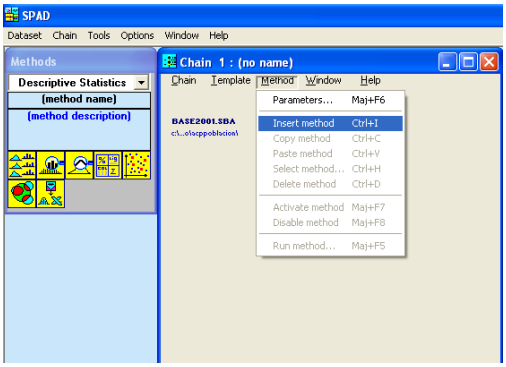


Figura 23

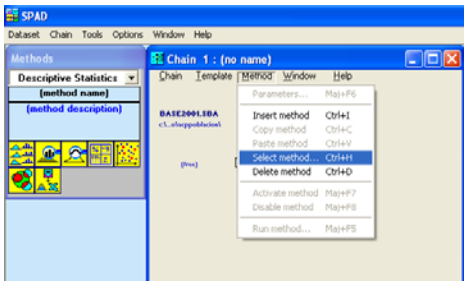


Figura 24

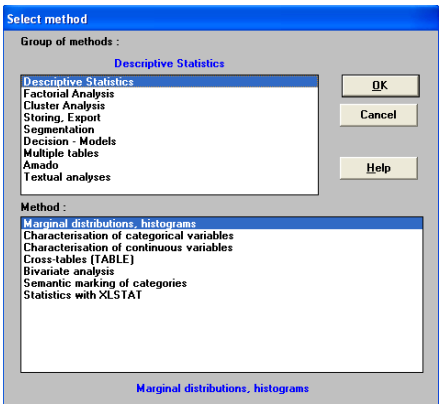


Figura 25

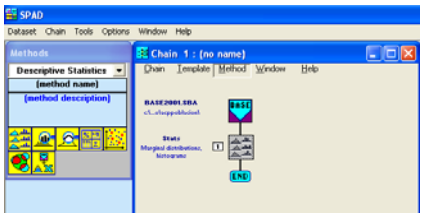


Figura 26

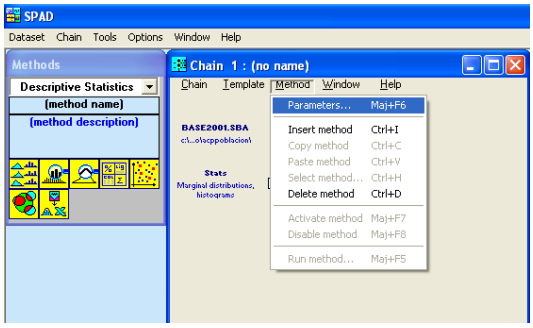


Figura 27

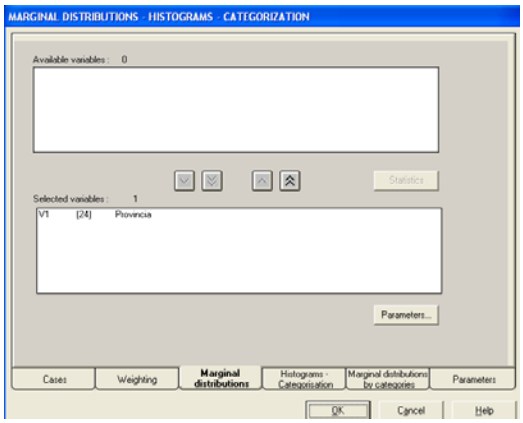


Figura 28

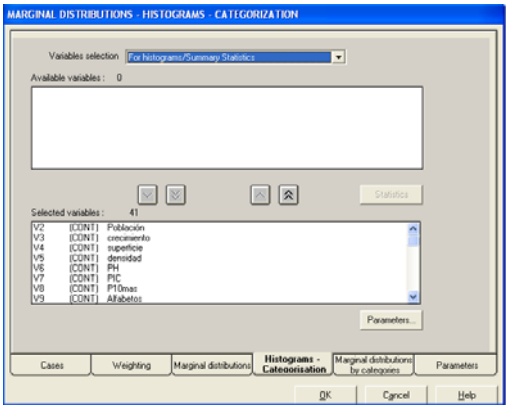


Figura 29

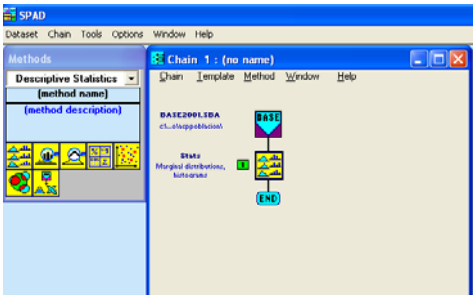


Figura 30

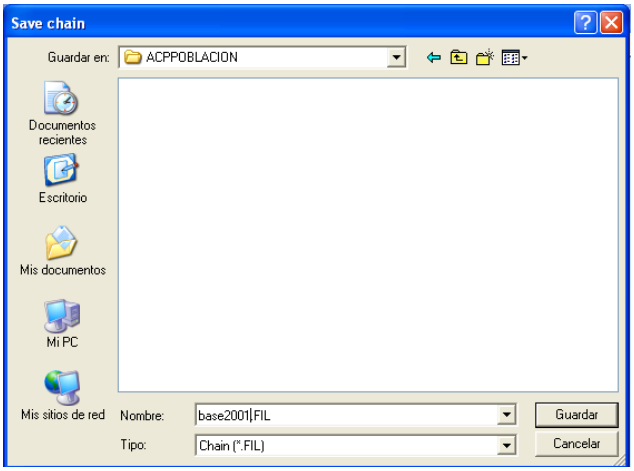


Figura 32

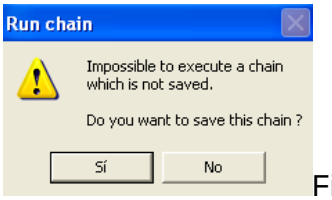


Figura 31

Figura 33

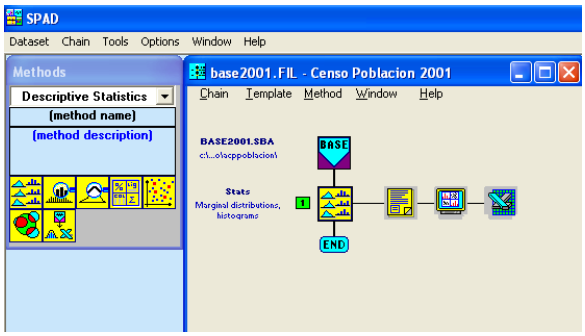


Figura 34

La ventana del entorno de trabajo, que inicialmente recibió el nombre de *Chain* y ahora tiene el nombre de *Base2001.fil*, tiene una columna con los procedimientos aplicados a partir de la cual se desprenden los resultados en fila.

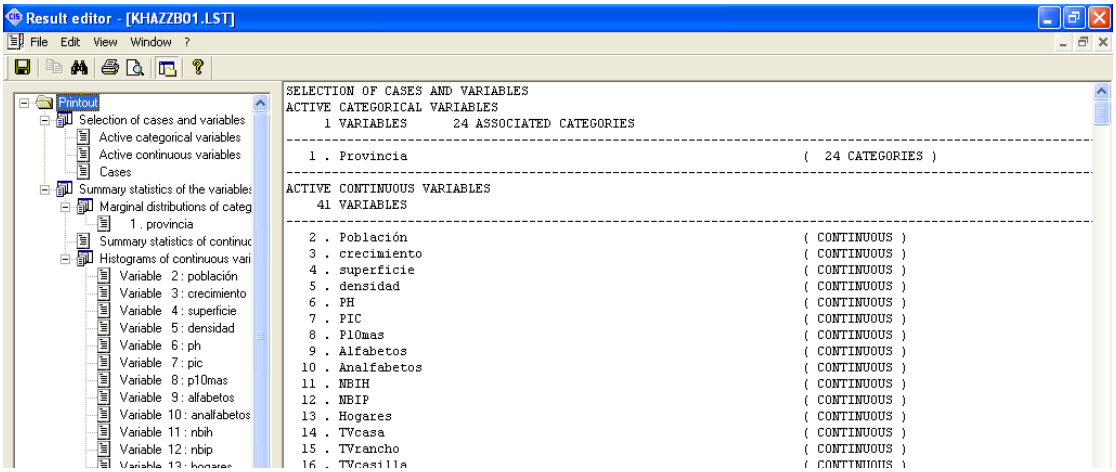


Figura 35

Análisis de Componentes Principales

Para realizar un nuevo análisis hay que situarse en el ícono desde donde se configuró el análisis anterior. Repetir el procedimiento de *Method-Insert method…* (Figura 36 y 37), *Method-Select method…* (Figura 38) seleccionar *Factorial Analysis* en la primer ventana y *Principal components analysis* en la segunda (Figuras 39 y 40), *Method Parameters* (Figura 41). Esta es la ventana donde se define sobre qué conjunto -de observaciones y variables- se realiza el análisis de componentes principales. En la solapa variables, al igual que en el caso anterior, se seleccionan las variables y se le asigna el rol de activas o suplementarias (ilustrativas).

Cada variable puede asumir un sólo tipo de rol y las cualitativas, en ACP, sólo pueden ser suplementarias, las cuantitativas pueden ser activas o suplementarias. Las variables activas son aquellas que participan en las operaciones matemáticas necesarias para arribar a un resultado en el marco del análisis definido; las variables ilustrativas no forman parte de este cálculo pero el método –y el software en buena medida- posibilita que ilustren el resultado final. Se observa que en esta solapa, además de las ventanas de variables disponibles y variables seleccionadas, una pestaña donde seleccionar el rol que va a asignarse a cada variable (Figuras 42, 43 y 44). En la solapa casos, por defecto están seleccionados todos, pero es posible seleccionar sólo algunos a través de una lista o un filtro lógico (Figura 45).

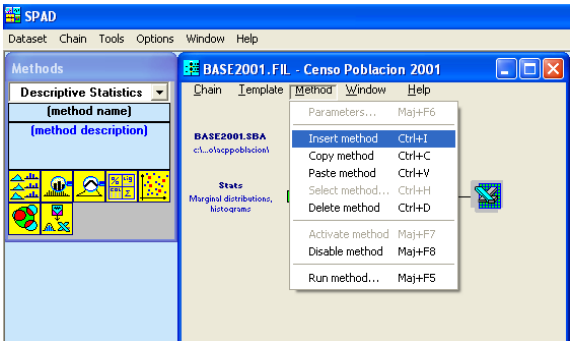


Figura 36

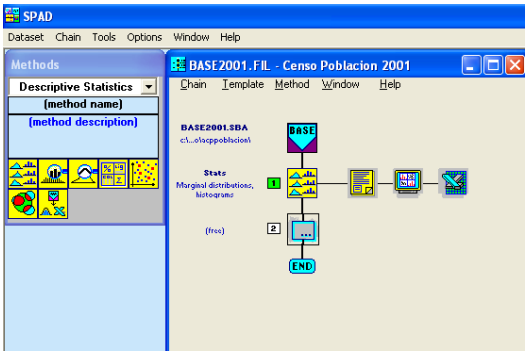


Figura 37

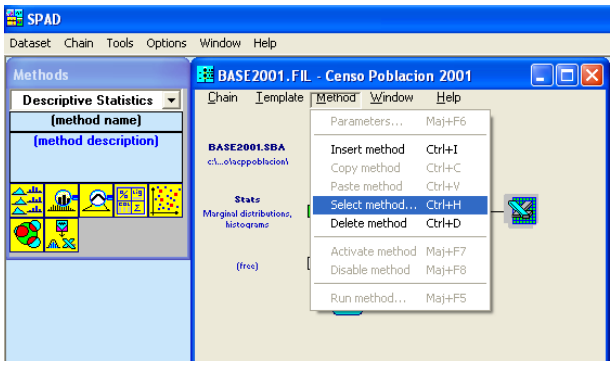


Figura 38

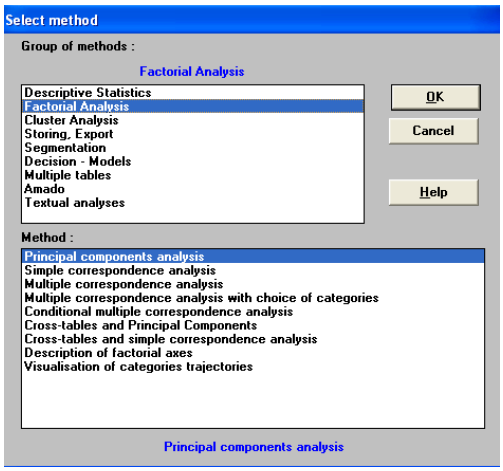


Figura 39

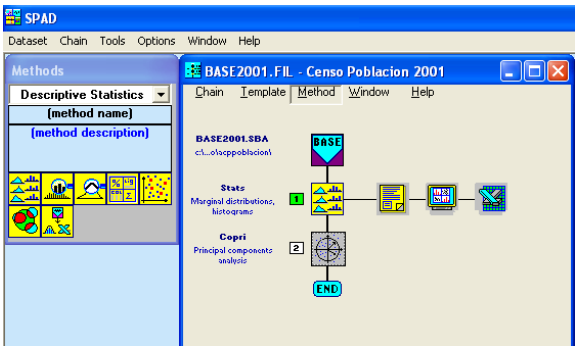


Figura 40

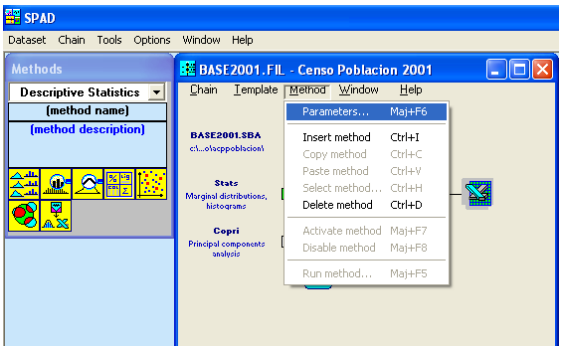


Figura 41

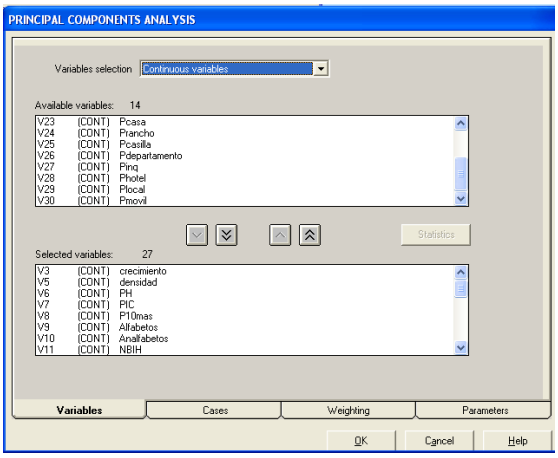


Figura 42

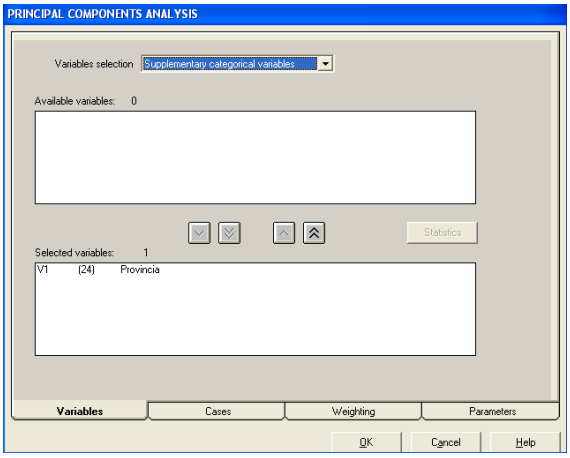


Figura 43

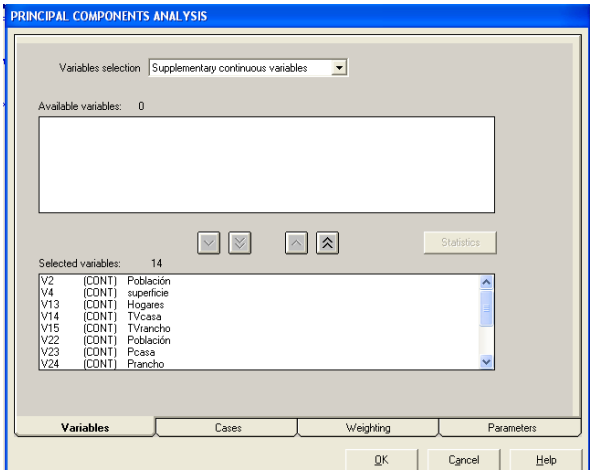


Figura 44

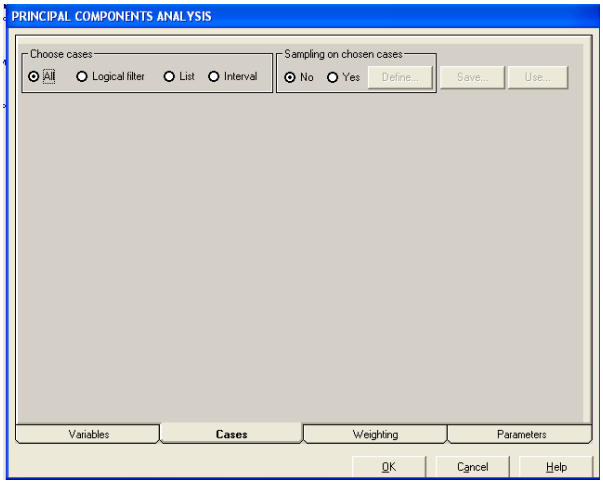


Figura 45

En la solapa *Weightin*, por defecto se tiene una ponderación Uniforme pero es posible asignar a las observaciones una ponderación diferente (Figura 46); en la solapa *Parameters* se puede pedir el resultado para todos los casos observados –el que será incluido dentro del informe general- y también en Excel desde el ícono *Options* (Figuras 47 y 48). Luego de dar todas estas indicaciones el ícono de ACP adquiere colores (Figura 49), por último queda por hacer *Method-Run* (Figura 50 y 51). Obsérvese que en este caso no es necesario generar un archivo *.fil, esto es así porque hay un análisis previo. Dicho de otra manera, el procedimiento descripto en las figuras 31 a 33 se hace para el primer

análisis de la base bajo estudio, todos los que siguen se acumulan en el mismo entorno de trabajo.

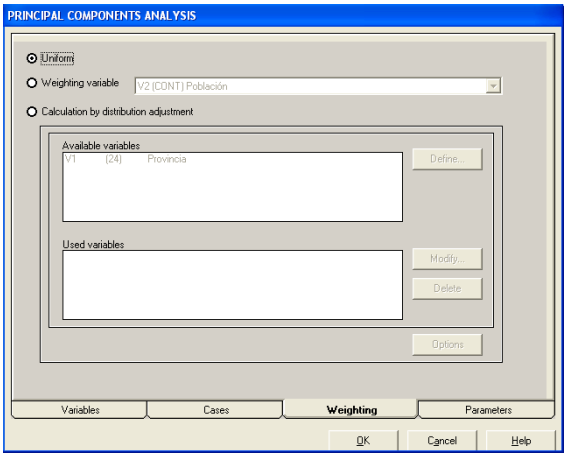


Figura 46

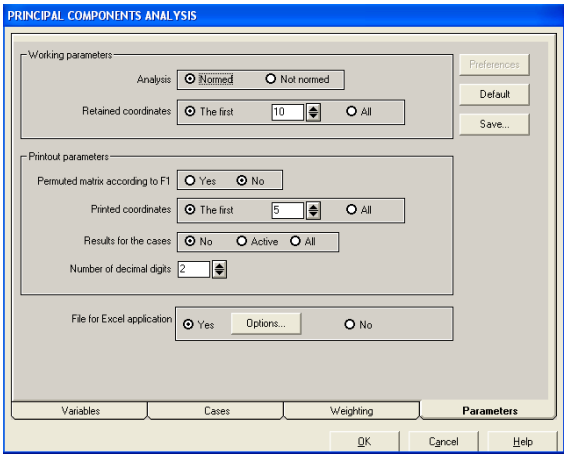


Figura 47

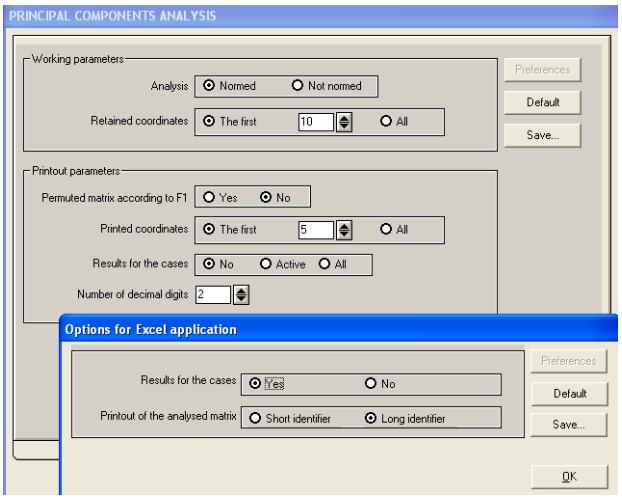


Figura 48

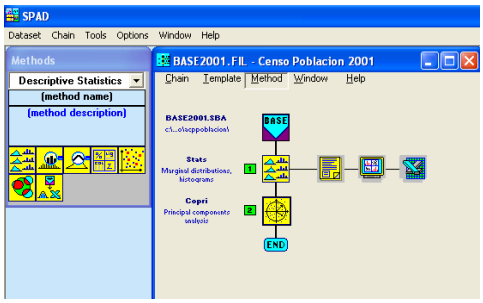


Figura 49

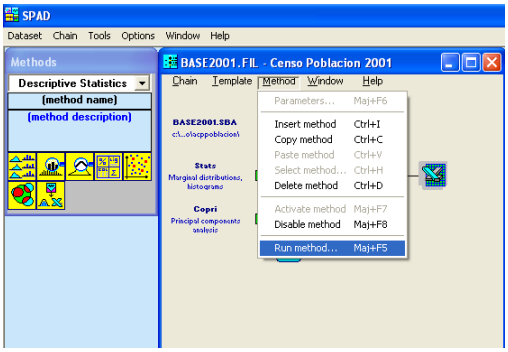


Figura 50

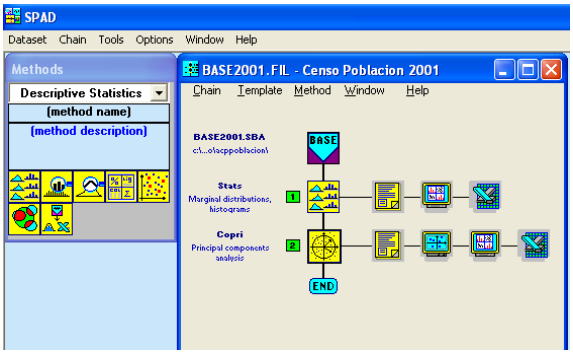


Figura 51

El ícono de texto en la salida de ACP contiene los indicadores que permiten tanto el análisis como la evaluación del método (Figura 52). Por ejemplo, puede verse en *Loadings of variables on axes 1 to 5* la oposición de las variables en cada eje. El eje 1 opone infraestructura física y personal, en el semieje negativo se tienen la población con 10 años o más, alfabetizada, en hogares con piso de calidad 1, calidad de materiales 1 con inodoros con descarga de agua y desagüe a cloacas en tipo de vivienda departamento; mientras que, en el semieje positivo se encuentran los analfabetos con NBI en hogares y en la población, que cuentan en los hogares con inodoro sin descarga de agua y calidad de pisos 2 y 3. El eje 2 opone las características de la infraestructura física a la que están expuestos los hogares y la población. El semieje negativo se caracteriza por la calidad de materiales para hogares que tienen inodoros con descarga de agua pero sin red cloacal, respecto al tipo de vivienda, es más frecuente la presencia de los tipos departamento y hotel con alta calidad de materiales e inodoros con descarga de agua a red pública. (Figura 53).

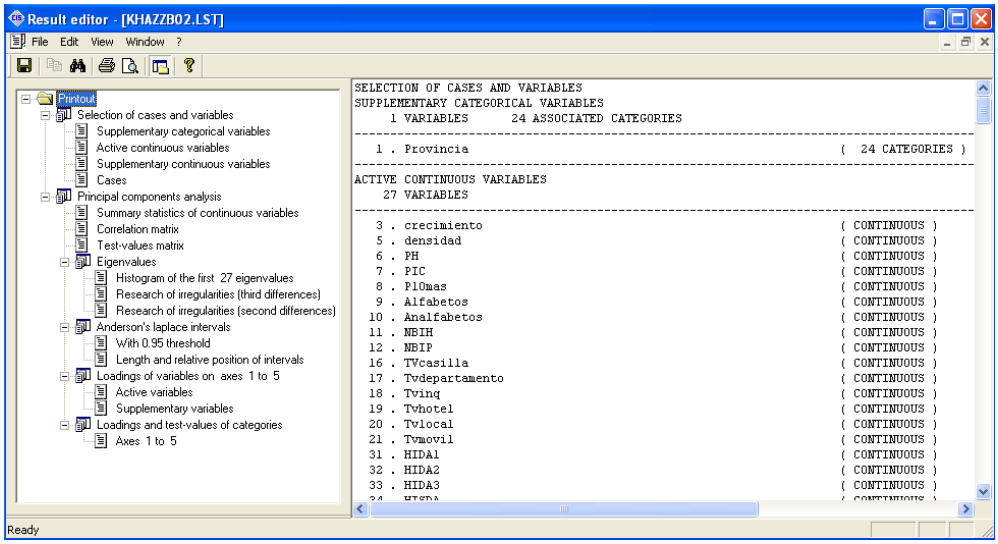


Figura 52

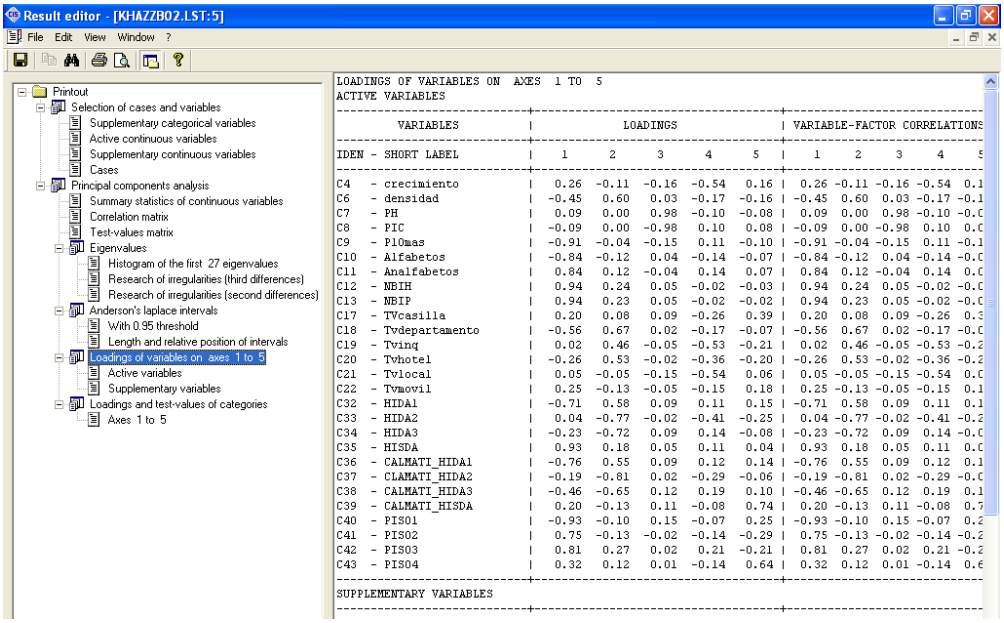


Figura 53

Desde *Loadings and test values of categories*, se observa la oposición en los ejes factoriales de las variables cualitativas. En este caso particular, para el primer factor se tiene, en el semieje positivo a las provincias de Chaco, Salta y Santiago del Estero y en el semieje negativo la provincia de Buenos Aires y la ciudad de Buenos Aires. En el segundo eje factorial, en el semieje positivo la ciudad de Buenos Aires y en el semieje negativo a las provincias de Buenos Aires y La Pampa (Figura 54).

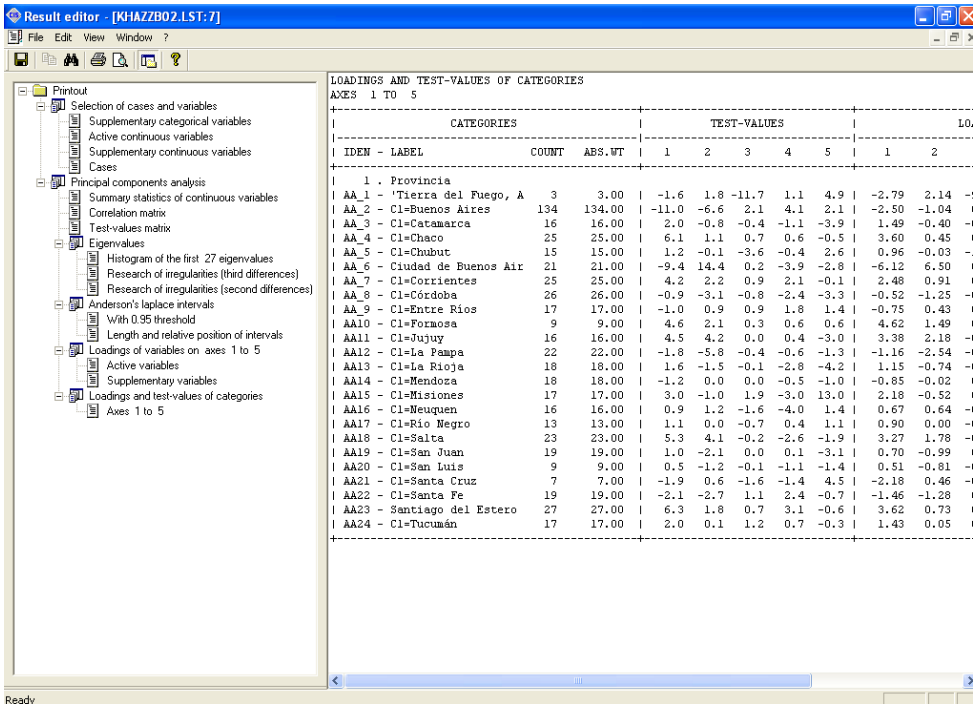


Figura 54

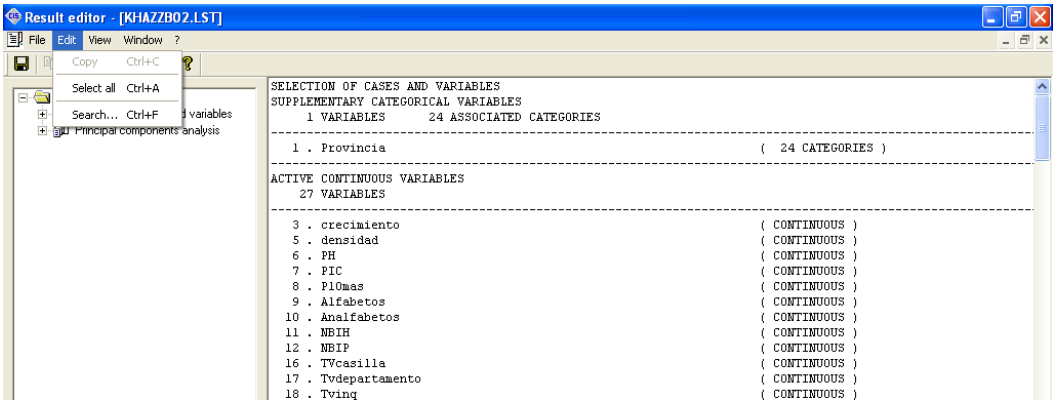


Figura 55

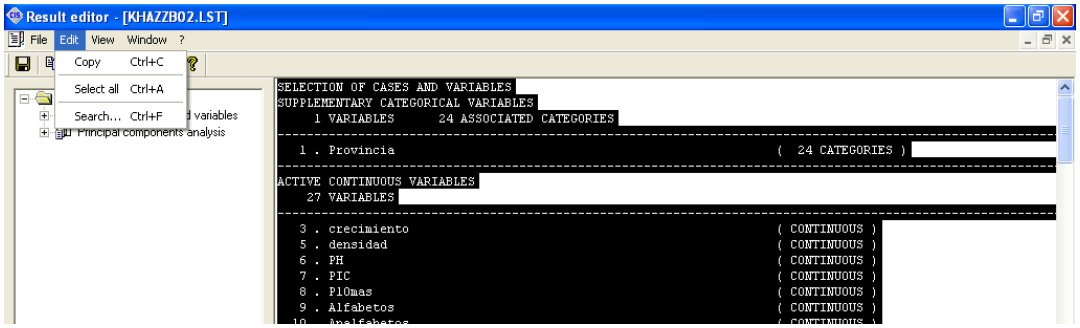


Figura 56

Esta salida puede llevarse a un editor de textos tipo Word desde *Edit-Select All* (Figura 55) y luego *Edit-Copy* (Figura 56). El resultado completo para el análisis realizado puede observarse en el Anexo Análisis de Componentes Principales.

Para graficar la dispersión de la nube de puntos es necesario situarse en el ícono gráfico (Figura 57), con doble click se accede al editor gráfico (Figura 58). Haciendo *Graph-New* se accede al cuadro de selección de elementos que definen el gráfico (Figura 59) para dar paso al mismo (Figura 60).

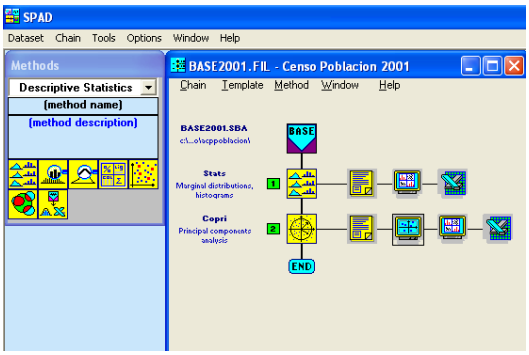


Figura 57

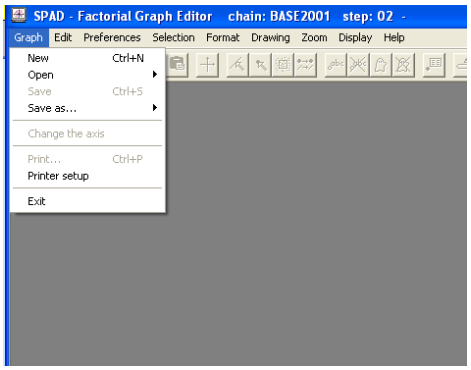


Figura 58

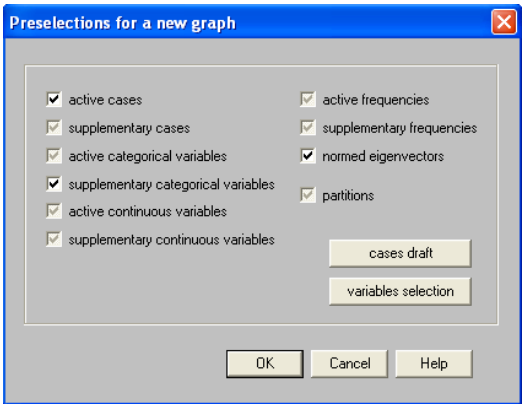


Figura 59

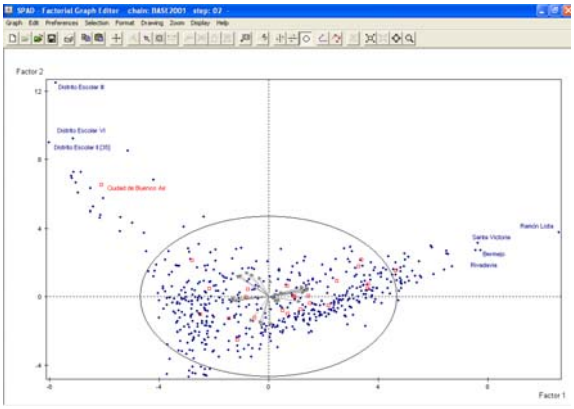


Figura 60

Clasificación de la nube de puntos en ACP

En este paso el procedimiento es agrupar las observaciones de a pares según su grado de similitud; el método consiste en medir las distancias entre las unidades de observación, aquel par que presente menor distancia da lugar a un nodo (unidad de observación virtual) la cual va a agruparse a aquella observación o nodo que se encuentre a menor distancia. Este es el paso previo y necesario para particionar la nube y hallar la composición de los grupos y su caracterización.

Este procedimiento requiere que previamente se haya realizado un análisis de componentes principales (ACP) o análisis factorial de correspondencias (AFC). Desde el ícono donde se configuró el ACP se sigue la secuencia *Method-Insert method…* (Figura 61), *Method-Select method…* (Figura 62), *Cluster Analysis* en la primer ventana y *Factor based cluster analysis* en la segunda ventana (Figura 63). En la columna de procedimientos se ve el nuevo ícono en gris (Figura 64), el paso siguiente es parametrizar el procedimiento. Desde *Method-Parameters…* se accede a la ventana de configuración por defecto ofrecida por SPAD a la que no es necesario hacer modificaciones (Figura 65). Al aceptar esta ventana el ícono adquiere color, lo cual significa que se han realizado los pasos convenientes para que pueda ejecutarse el método (Figura 66), haciendo *Method-Run…* (Figura 67) se genera la fila de resultados (Figura 68).

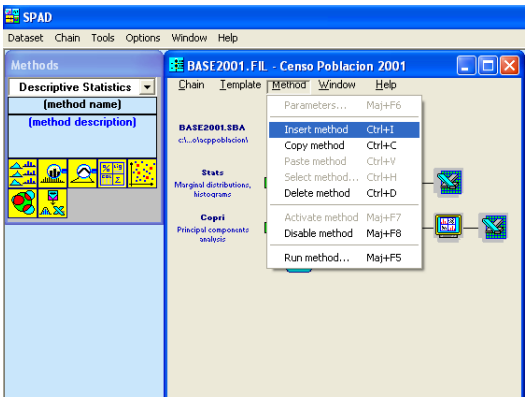


Figura 61

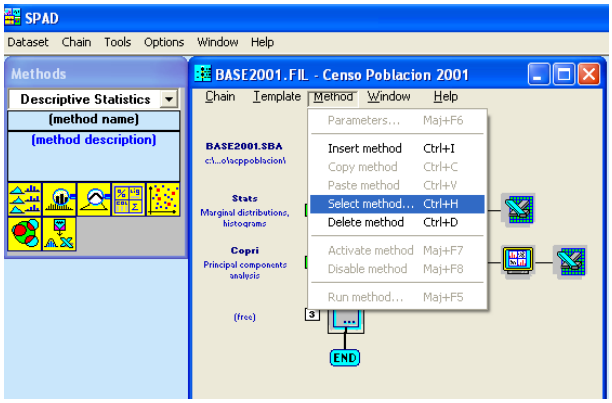


Figura 62

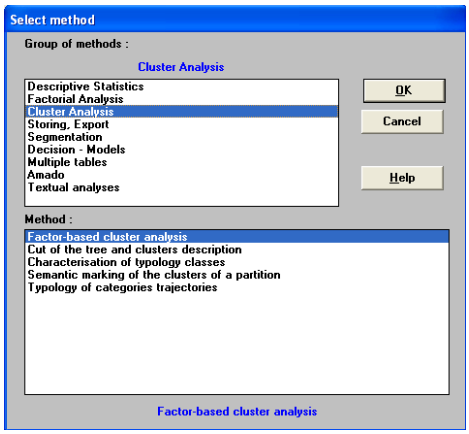


Figura 63

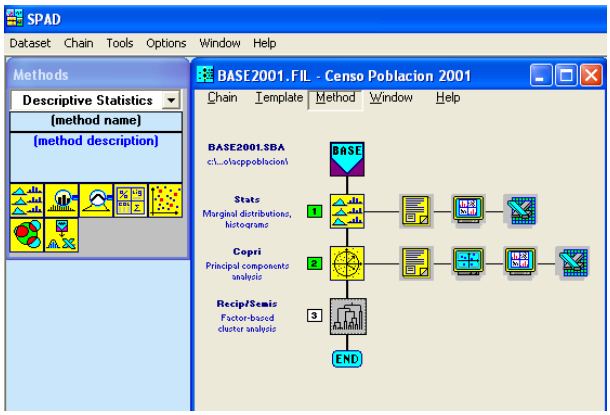


Figura 64

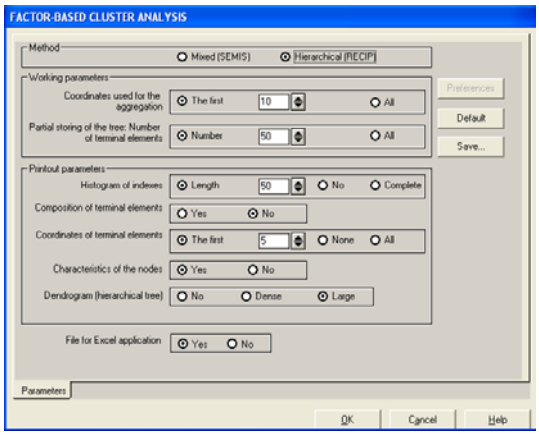


Figura 65

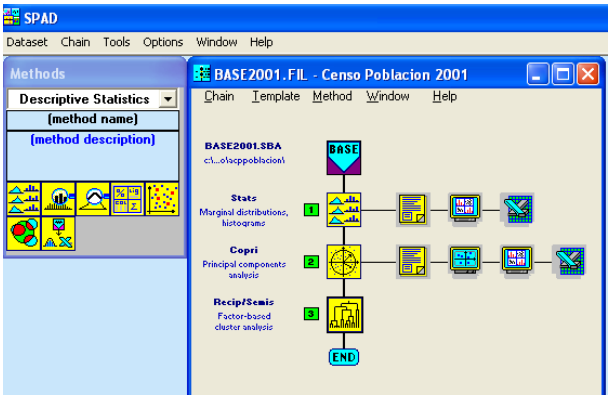


Figura 66

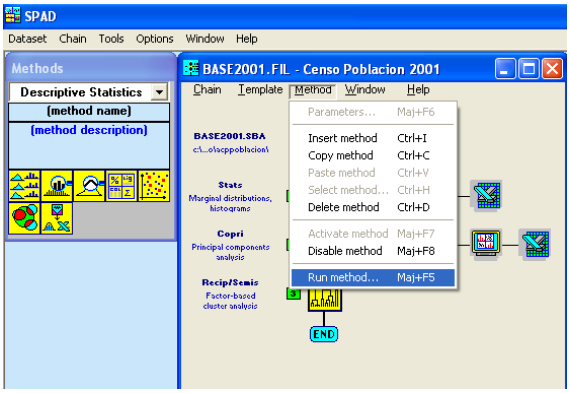


Figura 67

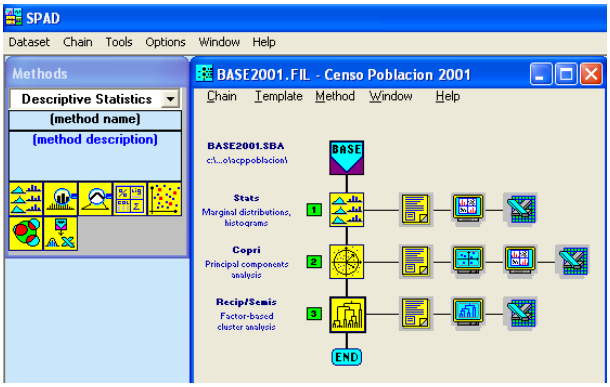


Figura 68

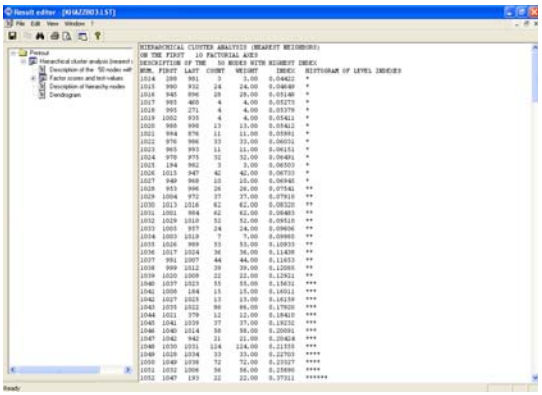


Figura 69

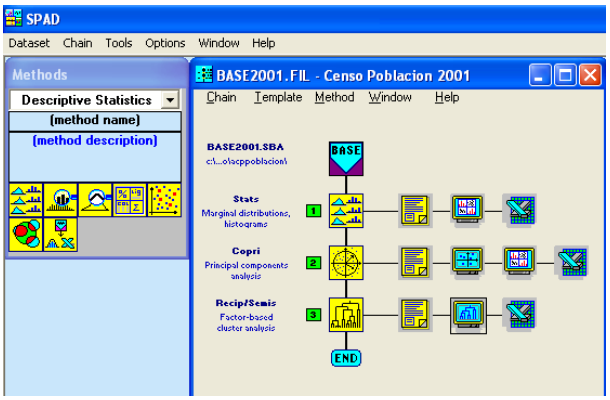


Figura 70

En la fila de resultados se tiene, como en los procedimientos anteriores, un ícono de texto, un ícono gráfico y un ícono de planilla de cálculo. En el de texto se informa cómo se han agregado los últimos nodos, los valores test correspondientes y la gráfica del dendograma (Figura 69). En el ícono gráfico se tiene el dendograma (Figura 70 y 71), la altura de las barras es indicativa de la cantidad de grupos que se pueden obtener en la partición.

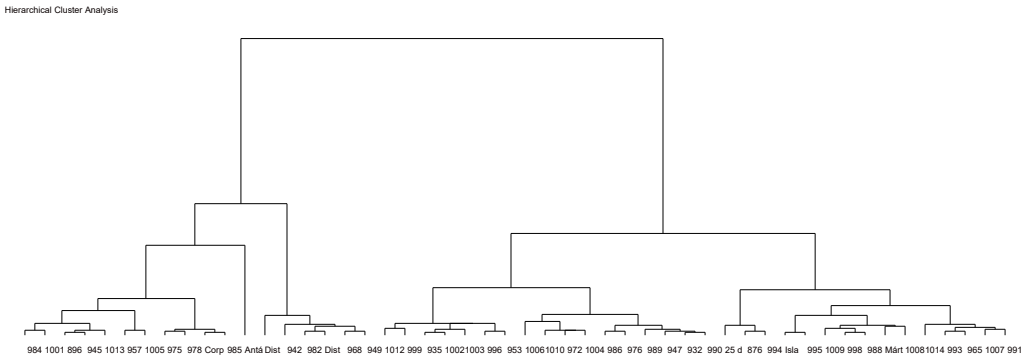


Figura 71

Partición de la nube de puntos

El paso siguiente es generar los grupos y describir cómo se compone cada uno de ellos. Para esto, desde el ícono donde se configuró la clasificación se sigue la secuencia *Method-Insert method…* (Figuras 72, 73 y 74), *Method-Select method….* (Figura 75), *Cluster Analysis* en la primer ventana y *Cut of the tree and clusters description* en la segunda ventana (Figuras 76 y 77)

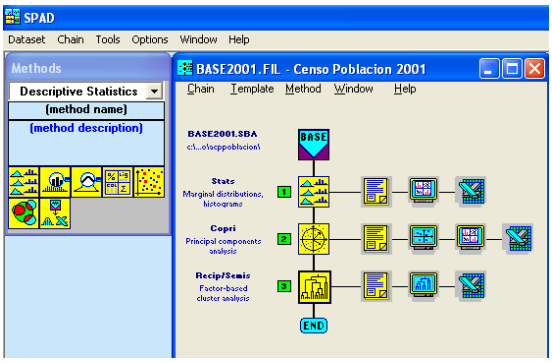


Figura 72

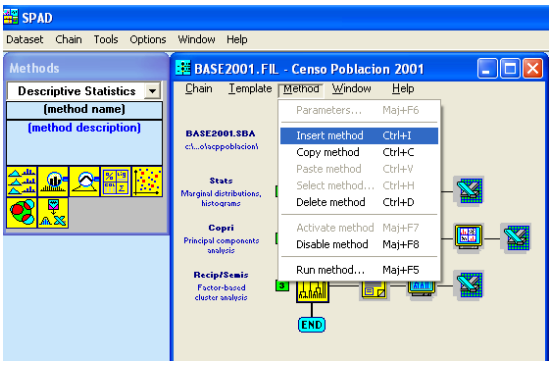


Figura 73

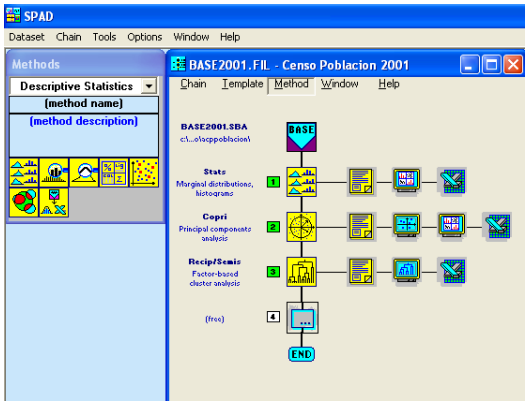


Figura 74

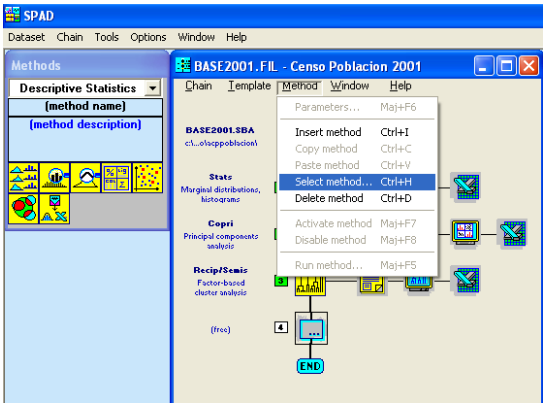
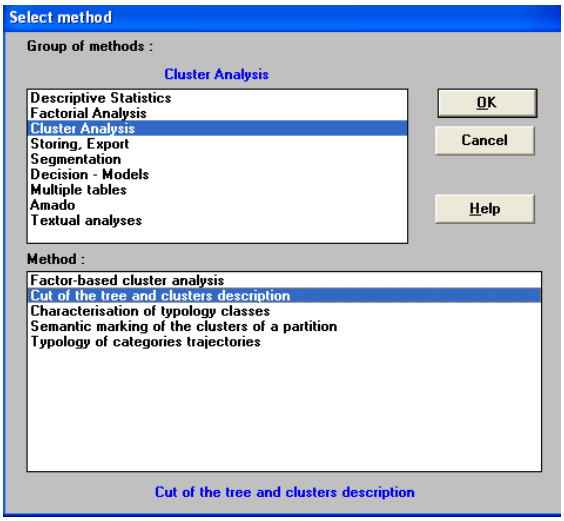


Figura 75



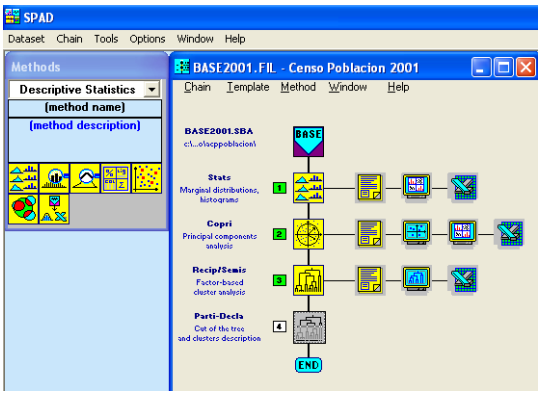


Figura 77

Figura 76

Para configurar la partición se accede desde *Method-Parameters….* en la primer solapa se deben elegir las dimensiones de la partición (el número de grupos), por defecto SPAD brindará 3 particiones de 3 a 10 grupos cada una (Figura 78); inicialmente es conveniente aceptar este procedimiento, luego de analizar la salida es posible pedirle que realice una partición con determinada cantidad de grupos. En la segunda solapa (*Partitioning parameters*) es oportuno, por las unidades de observación que se analizan, solicitarle que asigne cada caso al grupo de pertenencia, esto se logra desde *Printout parameters-CaseCluster correspondence-For all cases*, el resto está

dado por defecto y arroja los indicadores necesarios (Figura 79). En la tercer solapa (*Partitions characterisation)* es para ampliar el nivel de confianza de los indicadores, inicialmente se deja por defecto (Figura 80). Al aceptar la ventana de configuración, el ícono adquiere color, el paso siguiente es *Method-Run* para ejecutar el procedimiento (Figura 82) que da por resultado la fila de igual manera que en los casos anteriores (Figura 83).

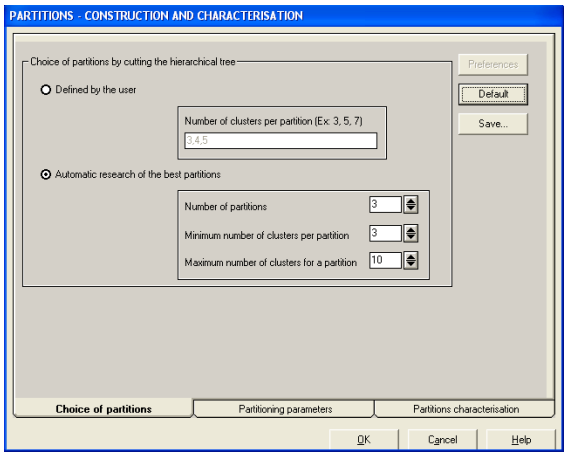


Figura 78

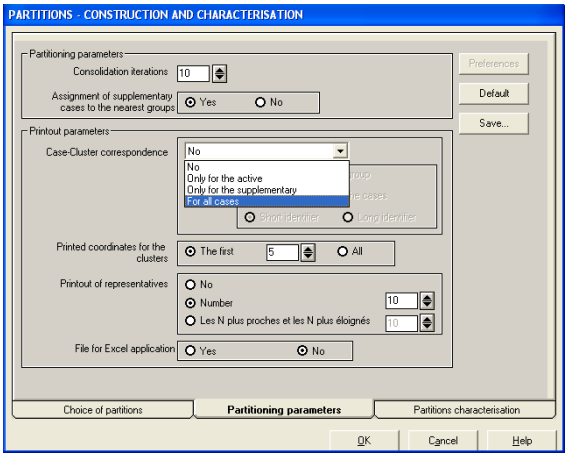


Figura 79

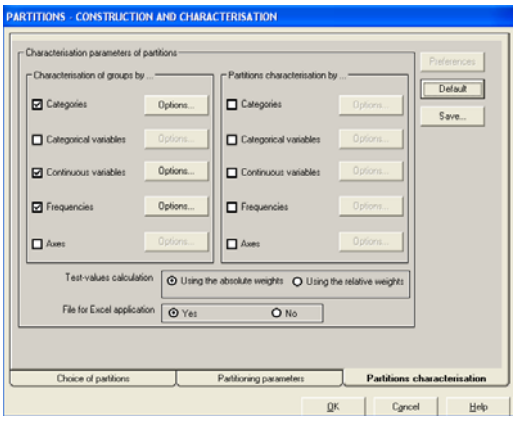


Figura 80

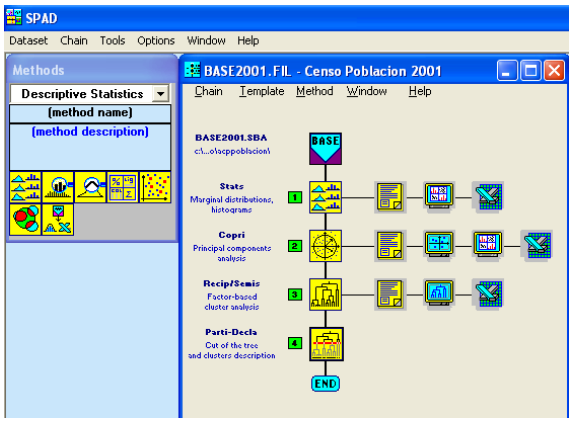


Figura 81

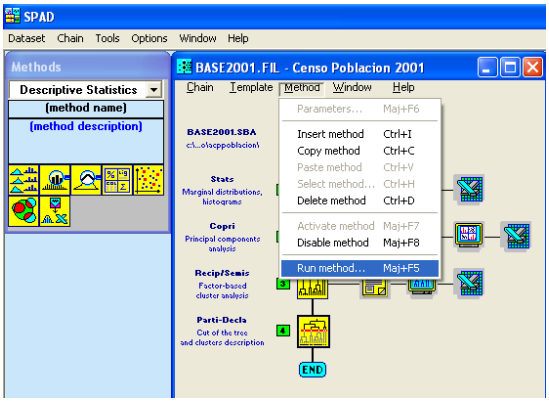


Figura 82

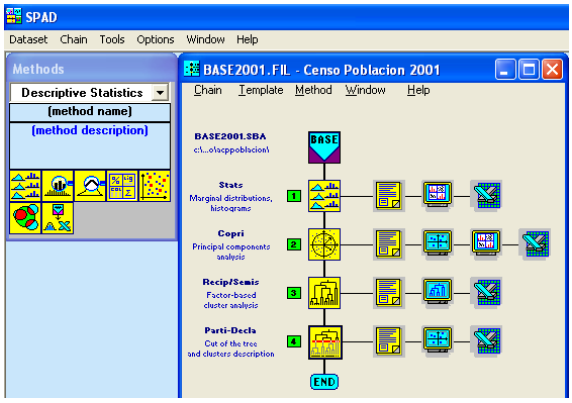


Figura 83

Nuevamente, el ícono texto tiene el resultado del procedimiento. En la primer pantalla informa los cortes que realizó en el árbol de clasificación; se observa que indica el corte de 5 y el de 10, en primer lugar siempre informa el que tiene mejores indicadores estadísticos (Figura 84); en *Clusters representatives* se tienen las unidades de observación más representativas de cada grupo (Figura 85); en *Description and characterisation of partitions* se describen las características de cada grupo, tanto para las variables ilustrativas como activas (Figuras 86 y 87). En el Anexo Partición de la nube de puntos se tiene la salida completa.

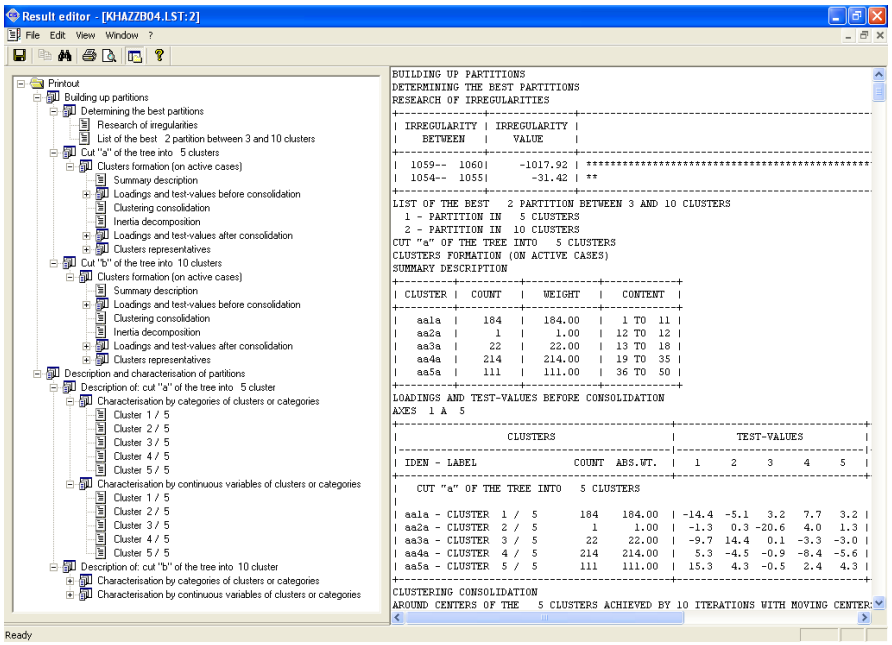


Figura 84

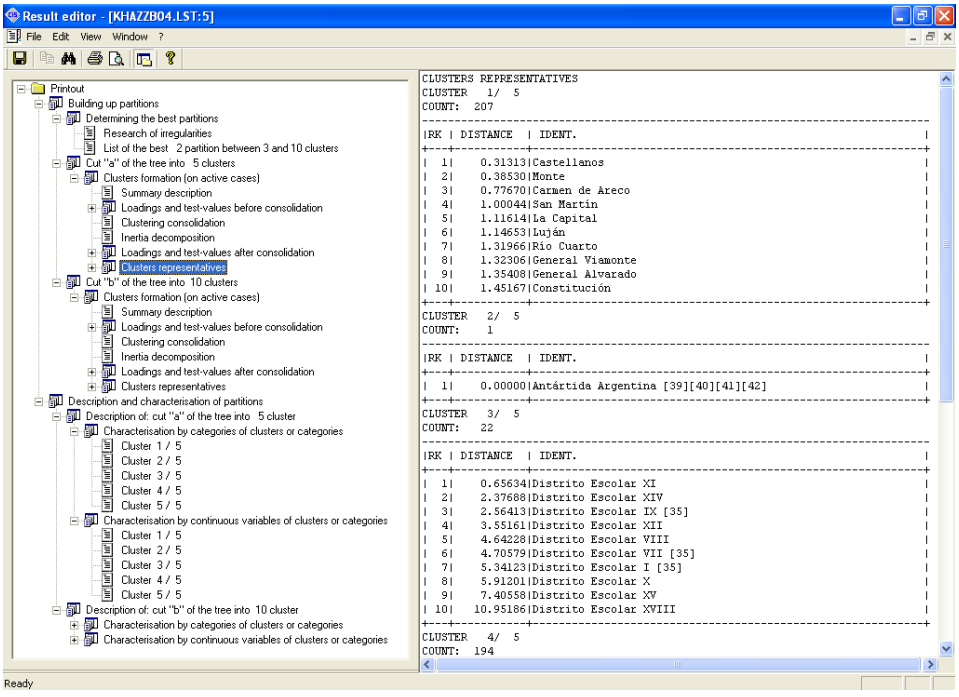


Figura 85

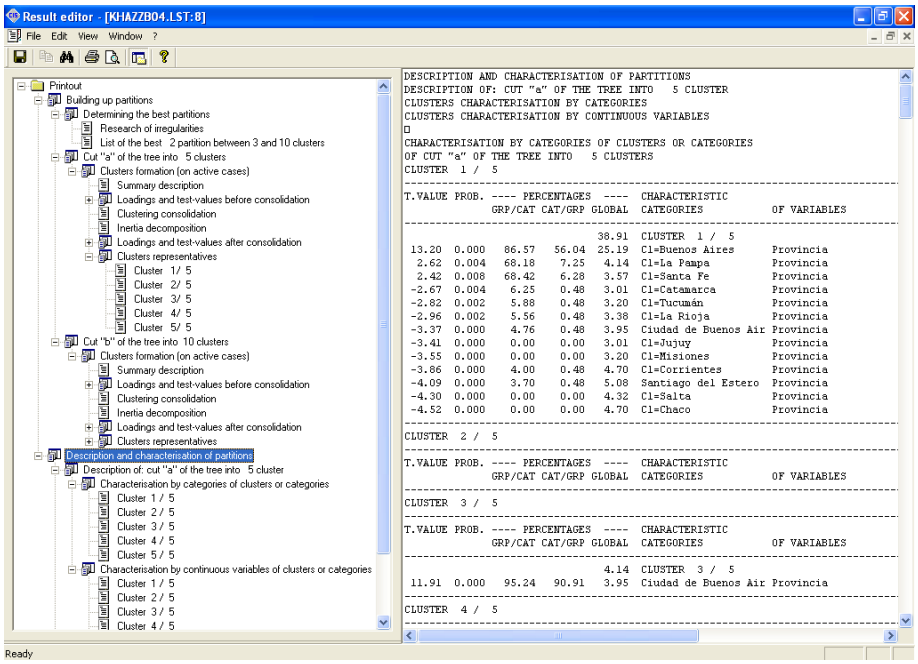


Figura 86

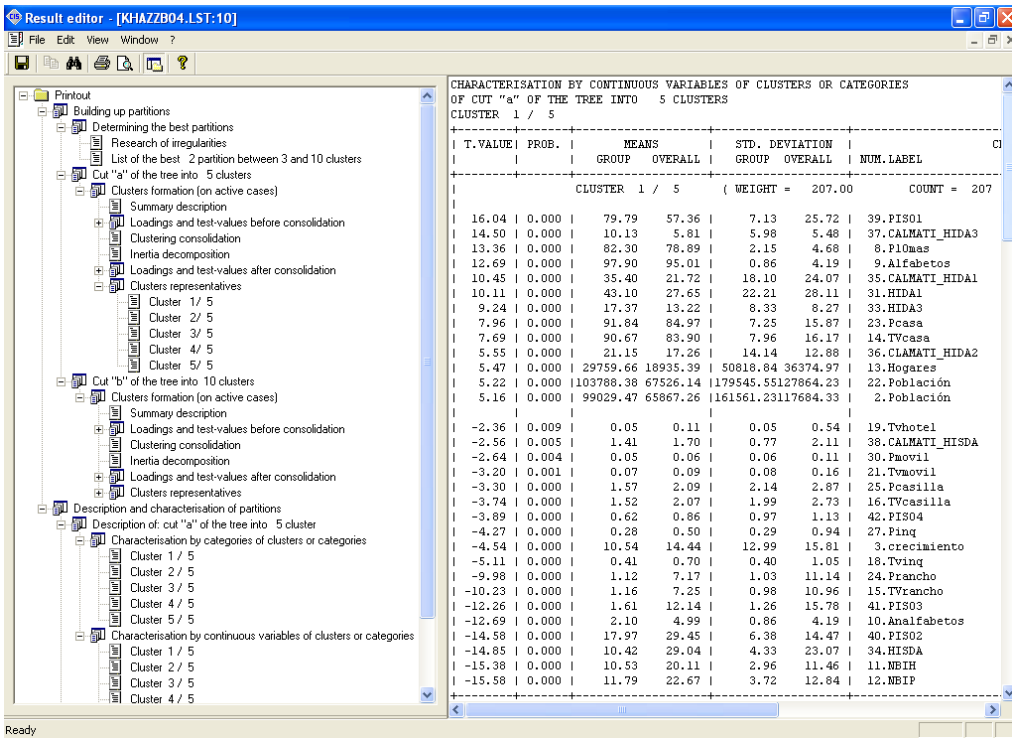


Figura 87

CASOS DE ESTUDIO, PREGUNTAS Y PROBLEMAS

Caso 1: Visitar páginas Web de instituciones de predicción.

Hay instituciones que realizan predicciones en el campo de la economía y de la gestión de empresas, cada una de las cuales está elaborada con criterios, planteamientos metodológicos y técnicas de predicción distintas. Además, la interpretación que cada institución hace de la realidad económica que le rodea es bien diferente, puesto que está basada en sus propias percepciones. Por ello, con frecuencia las predicciones realizadas sobre la marcha de la economía son desiguales. Sin embargo, todas ellas son de utilidad para el analista económico o empresarial.

Como actividad se le sugiere que realice una visita a las páginas Web de algunas de las instituciones que realizan predicciones, para alcanzar un doble objetivo:

En primer lugar, es probable que, desde el punto de vista de su propia práctica profesional, en algún momento deba hacerse una idea del entorno económico actual y las perspectivas que sobre el mismo tengan las diversas instituciones. Para ello deberá consultar los Informes de Predicción del mayor número posible de instituciones. Hoy día son varios los centros de predicción que proveen esta información gratuitamente a través de Internet. Le sugerimos que visite las páginas Web de algunas de estas Instituciones. Localice el lugar concreto de la página en el que la Institución coloca sus informes de coyuntura y predicción (le sugerimos que agregue las direcciones en su carpeta de favoritos), y deténgase a analizar con detalle la estructura de los distintos informes: tipo de información, variables sobre las que se

realiza predicción, forma de estructurarla... Observe también las diferencias entre las distintas instituciones en cuanto a la forma de presentar la información. Analice también los comentarios que los autores de los informes realizan a partir de la predicción, puesto que en su categoría de expertos seguidores de la realidad económica sus comentarios dan pistas interesantes sobre las perspectivas de la economía.

FONDO MONETARIO INTERNACIONAL, World Economic Outlook http://www.imf.org/external/pubind.htm	Informes semestrales, normalmente en abril y noviembre. Predicción con gran desagregación de variables, países y grandes áreas geográficas. Incluye además análisis de temas de actualidad.
COMISIÓN EUROPEA, European Economy http://europa.eu.int/comm/economy_finance/index_en.htm	Informes semestrales en primavera y otoño. Predicciones para los países de la Unión, con amplia desagregación. Incluye también una buena base histórica.
MINISTERIO DE ECONOMÍA, Previsiones Macroeconómicas; The Spanish Economy; y Actualización del Programa de Estabilidad http://www.mineco.es/sgpc/405SGPCM.htm	Predicciones oficiales, pero sin actualización periódica determinada. Interesante porque muestra la visión gubernamental de la marcha de la economía.
ORGANIZACIÓN PARA LA COOPERACIÓN Y EL DESARROLLO ECONÓMICO, OECD Economic Outlook http://www.oecd.org/publications/	Amplia desagregación por países e indicadores. Con análisis de actualidad
INSTITUTO FLORES DE LEMUS DE ESTUDIOS AVANZADOS DE ECONOMIA (UNIVERSIDAD CARLOS III), Boletín Inflación y Análisis Macroeconómico http://www.uc3m.es/uc3m/inst/FL/IFL.HTM http://www.uc3m.es/uc3m/inst/FL/boletin/index.html.	Predicción de un número reducido de variables. Predicciones limitadas a suscritos. También se puede acceder al informe de la red European Forecasting Network sobre la economía europea, en el que también participa el grupo AQR de la Universidad de Barcelona.
BBVA, Situación España http://www.bbva.es (Servicio de Estudios/situación)	Informe mensual de cierta tradición. Incluye informe de coyuntura, predicciones de un cuadro básico y notas de actualidad.
COMMERZBANK, Economic Research https://www.commerzbank.com/ (Research/Economic Research)	Predicciones sobre los principales indicadores económicos de varios países, incluido España
FUNCAS, Previsiones FUNCAS; Panel de Previsiones http://www.funcas.ceca.es/ (Indicadores/Coyuntura Nacional e internacional)	Se incluye predicción sobre un cuadro macroeconómico de actualización bimensual básico. El panel incluye una síntesis de las predicciones de varias instituciones

Figura 1. Algunas referencias de informes de predicción

En segundo lugar, y como complemento a lo anterior, le sugerimos que elabore un breve informe de situación y perspectivas a partir de la información contenida en los informes de estas instituciones. Realice su propia clasificación de la información, sus propias tablas agrupando los datos, realice tablas de discrepancias entre instituciones y, por último, sintetice la información con unos breves comentarios. Con todo ello, podrá tener una visión de conjunto de la situación económica actual y su propio informe de predicción.

Caso 2 Localización de información y análisis de indicadores adelantados.

A lo largo del cuaderno, se ha puesto de manifiesto la importancia de la disponibilidad de información estadística y se han descripto los indicadores adelantados como una de las técnicas que se utilizan para llevar a cabo la predicción de una variable cuando no se dispone de información histórica de la misma, o se renuncia a la misma. En algunos casos, existen magnitudes económicas estrechamente relacionadas con otras, de las que sí es posible disponer de información estadística o de mejor calidad: frecuencia, actualización, disponibilidad electrónica.

Además, algunos indicadores presentan una característica clave a efectos de predicción; y es que algunos de ellos presentan la propiedad de “adelantar” el comportamiento de algunas variables que describen el comportamiento de algunos sectores. Por ejemplo, la evidencia muestra que una caída hoy en el indicador conocido como consumo aparente de cemento, anticipa una contracción en magnitudes que describen la actividad económica global del sector de la construcción, como es el Valor Añadido Bruto. Téngase en cuenta, que este sector representa una parte muy significativa de la economía de un país y el conocimiento sobre su comportamiento, puede ayudar en la anticipación sobre la evolución de magnitudes más agregadas, como el PIB.

Para tomar conocimiento de los principales indicadores que describen de manera desagregada el funcionamiento de una economía, se le sugiere que realice un breve informe sobre los indicadores de un sector. Ello, además, le permitirá ganar soltura para la obtención de la información estadística necesaria. Elija el sector de la economía en el que esté más interesado (industria, construcción, energía,…). Para tal sector, seleccione los indicadores que describen la actividad económica del mismo. Para ello, le sugerimos que revise anuarios estadísticos, informes patronales o páginas Web. Busque información que analice la posibilidad de que alguno de esos indicadores sea adelantado.

Por último le sugerimos que realice un análisis descriptivo de tales indicadores: analice las tasas de variación, su evolución histórica, busque coincidencias entre los mismos y, sobre todo, y en la medida de las disponibilidades estadísticas, analice las coincidencias con la magnitud. Aunque no sea más que de un modo intuitivo ¿se atrevería a decir que son indicadores adelantados? ¿Se arriesgaría a realizar con ellos una predicción de la magnitud?

Caso 3: Cálculo de Medias móviles

Una empresa decide aplicar la técnica de medias móviles como método rápido y sencillo de predicción de ventas, a efectos de determinar las necesidades de personal temporal de ventas en un período normal (sin oscilaciones estacionales ni tendencia, al menos especialmente significativas). A partir de la serie *y* (ventas semanales en millones de pesos), se pide calcular en excel:

- Medias móviles asimétricas de tres términos (M3t).
- Medias móviles asimétricas de seis términos (M6t).
- Medias móviles asimétricas ponderadas de tres términos y ponderaciones 3, 2, 1 (M3wt).
- Doble media móvil de tres y tres términos (M3.3t).

Además realice predicción para dos períodos. Utilice para todo ello la información de la Figura 3. Para trabajar copie esta tabla en un libro Excel (observe que el primer dato esté ubicado en la **CELDA B17 Y SE REFIERA AL PERÍODO 1** y siga la solución para cada apartado que se detalla a continuación).

	A	B	C	D	E	F	G
16		Periodo	VENTAS	M3t	M6t	M3wt	M3.3t
17		1	5,3	--	--	--	--
18		2	4,4	--	--	--	--
19		3	5,4		--		--
20		4	5,8		--		--
21		5	5,6		--		
22		6	4,8				
23		7	5,6				
24		8	5,6				
25		9	5,4				
26		10	6,5				
27		11	5,1				
28		12	5,8				
29		13	5,0				
30		14	6,2				
31		15	5,6				
32		16	6,7				
33		17	5,2				
34		18	5,5				
35		19	5,8				
36		20	5,1				
37		21	5,8				
38		22	6,7				
39		23	5,2				
40		24	6,0				
41		25	5,8				
42	Predicción	26	--				
43		27	--				

Figura 2 Ventas semanales

a) MEDIAS MÓVILES DE TRES TÉRMINOS.

La media móvil de tres términos se calcularía asignando a cada momento del tiempo (t) el resultado de sumar los valores de serie *y* en el momento *t*, *t*-1, y *t*-2, y dividir por 3. Por ejemplo, la media móvil de orden 3 para el período 10 se calcularía como (6,5+5,4+5,6)/3 = 5,83.

Para el cálculo en Excel sitúese en la celda D19 y escriba en lenguaje Excel la fórmula de la media móvil de ese período de tiempo, es decir, =(C19+C18+C17)/3 y pulse INTRO. Vuelva a marcar la celda D19 y sitúe el cursor en la esquina inferior derecha, hasta que el cursor se transforme en forma de aspa. Haga CLICK y arrastre el cursor hasta D41. Excel generará automáticamente la media móvil para toda la serie.

Para realizar la predicción para el período 26, asumimos como predicción el último valor de la serie de medias móviles, es decir, la predicción de la serie *y* en 26 es igual a M3t en el período 25. La predicción en el período 27 exigiría calcular la media móvil **de** orden 3. Para ello utilizaremos el valor de predicción en el período 26, y los valores de la serie *y* en los períodos 24 y 25.

En lenguaje Excel, escribiríamos en la celda D42, la fórmula =D41, mientras que en la celda D43 escribiríamos =(D42+C41+C40)/3.

b) MEDIAS MÓVILES DE SEIS TÉRMINOS.

La media móvil de seis términos se calcula en forma análoga. Por ejemplo, la media móvil de orden 6 para el período 10 se calcularía como (6,5+5,4+5,6+5,6+4,8+5,6)/6 = 5,58.

En Excel la fórmula de la media móvil para el primer período posible, el 6, sería (C22+C21+C20+C19+C18+C17)/6. La predicción del

período 26 se formularía como =E41, y la del período 27 como =(E42+C41+C40+C39+C38+C37)/6.

c) MEDIAS MÓVILES PONDERADAS DE TRES TÉRMINOS.

El cálculo es semejante, pero ahora atribuimos a cada valor de la serie *y* la ponderación correspondiente. Por ejemplo, la media móvil de orden 3 para el período 10 se calcularía como el cociente entre el resultado de (6,5 x 3)+(5,4 x 2)+(5,6 x 1), y la suma de las ponderaciones, es decir (3+2+1). Por tanto el valor para el período 10 sería igual a (35,9/6)=5,98. En lenguaje Excel se anota en la celda F19 la fórmula =(3*C19+2*C18+1*C17)/(3+2+1), y arrastraríamos el resultado hasta la celda F41.

Para la predicción se actúa en forma análoga a las anteriores. La predicción para el período 26 sería la media móvil obtenida en el período 25 (=F41), mientras que para la predicción del período 27, se calcularía la medida móvil ponderada incluyendo el valor predicho para el período 26, =(3*F42+2*C41+1*C40)/(3+2+1).

d) DOBLE MEDIA MÓVIL DE TRES Y TRES TÉRMINOS.

Se distinguen dos cálculos: en el primero se calcula la media móvil simple de orden 3 (punto a). En el segundo cálculo, se elabora una media móvil de orden 3 a partir de la serie obtenida por media móvil simple. Por ejemplo, el valor asignado a la doble media móvil de orden 3, para el período 10 viene dada por los valores de la media móvil simple de orden 3 para el período 10, 9 y 8. En definitiva, se calcularía como (5,8+5,5+5,3)/3= 5,53.

Por tanto, en lenguaje Excel anotaríamos en la celda G20 la fórmula =(D21+D20+D19)/3, arrastrando hasta G41. La predicción para el período 26 se obtendría como =(G41+D41+D40)/3, mientras que la predicción del período 27 se obtiene a partir de la fórmula

=(G42+G41+D41)/3. En definitiva, los resultados que obtenga deben ser similares a los recogidos en la siguiente tabla de la hoja Resultados. Además se recoge el error cuadrático medio (ECM) y el porcentaje de error absoluto medio (PEAM) de cada variante de predicción.

En la Figura 3 se recogen los resultados. Además se calcula el error cuadrático medio (ECM) y el porcentaje de error absoluto medio (PEAM) de cada variante de predicción.

En la Figura 4 se ilustran los valores observados de la variable ventas y sus pronósticos, que refleja el modo en que estas variantes de predicción transforman la serie original. Obsérvese, en especial, el modo en que las técnicas "alisan" los valores originales, obteniendo una serie con menor variabilidad.

	Datos	Medias Móviles				Errores cuadráticos medios				Porcentajes medios de error absoluto			
Periodo	VENTAS	M3t	M6t	M3wt	M3.3t	ECM(M3t)	ECM(M6t)	ECM(M3wt)	ECM(M3.3t)	PEAM(M3t)	PEAM(M6t)	PEAM(M3wt)	PEAM(M3.3t)
1	5,3					---	---	---	---	---	---	---	---
2	4,4					---	---	---	---	---	---	---	---
3	5,4	5,0		5,1		0,13	---	0,12	---	6,79	---	6,48	---
4	5,8	5,2		5,4		0,36	---	0,13	---	10,34	---	6,32	---
5	5,6	5,6		5,6	5,3	0,00	---	0,00	0,10	0,00	---	0,60	5,75
6	4,8	5,4	5,2	5,2	5,4	0,36	0,17	0,19	0,36	12,50	8,68	9,03	12,50
7	5,6	5,3	5,3	5,3	5,4	0,07	0,11	0,07	0,02	4,76	5,95	4,76	2,78
8	5,6	5,3	5,5	5,5	5,4	0,07	0,02	0,02	0,06	4,76	2,38	2,38	4,37
9	5,4	5,5	5,5	5,5	5,4	0,02	0,00	0,01	0,00	2,47	1,23	1,85	0,00
10	6,5	5,8	5,6	6,0	5,6	0,44	0,84	0,27	0,87	10,26	14,10	7,95	14,36
11	5,1	5,7	5,5	5,6	5,7	0,32	0,16	0,27	0,33	11,11	7,84	10,13	11,33
12	5,8	5,8	5,7	5,7	5,8	0,00	0,02	0,01	0,00	0,00	2,30	2,01	0,57
13	5,0	5,3	5,6	5,3	5,6	0,09	0,32	0,08	0,35	6,00	11,33	5,67	11,78
14	6,2	5,7	5,7	5,7	5,6	0,28	0,28	0,22	0,37	8,60	8,60	7,53	9,86
15	5,6	5,6	5,7	5,7	5,5	0,00	0,01	0,01	0,01	0,00	1,79	1,79	1,39
16	6,7	6,2	5,7	6,3	5,8	0,28	0,93	0,20	0,79	7,96	14,43	6,72	13,27
17	5,2	5,8	5,8	5,8	5,9	0,40	0,30	0,32	0,44	12,18	10,58	10,90	12,82
18	5,5	5,8	5,7	5,6	5,9	0,09	0,04	0,01	0,19	5,45	3,64	1,82	7,88
19	5,8	5,5	5,8	5,6	5,7	0,09	0,00	0,04	0,01	5,17	0,57	3,45	1,53
20	5,1	5,5	5,7	5,4	5,6	0,13	0,30	0,09	0,24	7,19	10,78	5,88	9,59
21	5,8	5,6	5,7	5,6	5,5	0,05	0,01	0,05	0,08	4,02	2,01	4,02	4,98
22	6,7	5,9	5,7	6,1	5,6	0,69	1,03	0,32	1,14	12,44	15,17	8,46	15,92
23	5,2	5,9	5,7	5,8	5,8	0,49	0,23	0,36	0,33	13,46	9,29	11,54	11,11
24	6,0	6,0	5,8	5,9	5,9	0,00	0,05	0,02	0,01	0,56	3,89	2,50	1,48
25	5,8	5,7	5,8	5,8	5,8	0,02	0,00	0,00	0,00	2,30	0,57	0,57	0,77
Predicción 26		*5,7*	*5,8*	*5,8*	*5,8*	---	---	---	---	---	---	---	---
Predicción 27		*5,8*	*5,9*	*5,8*	*5,8*	---	---	---	---	---	---	---	---

Figura 4 Pronóstico de ventas

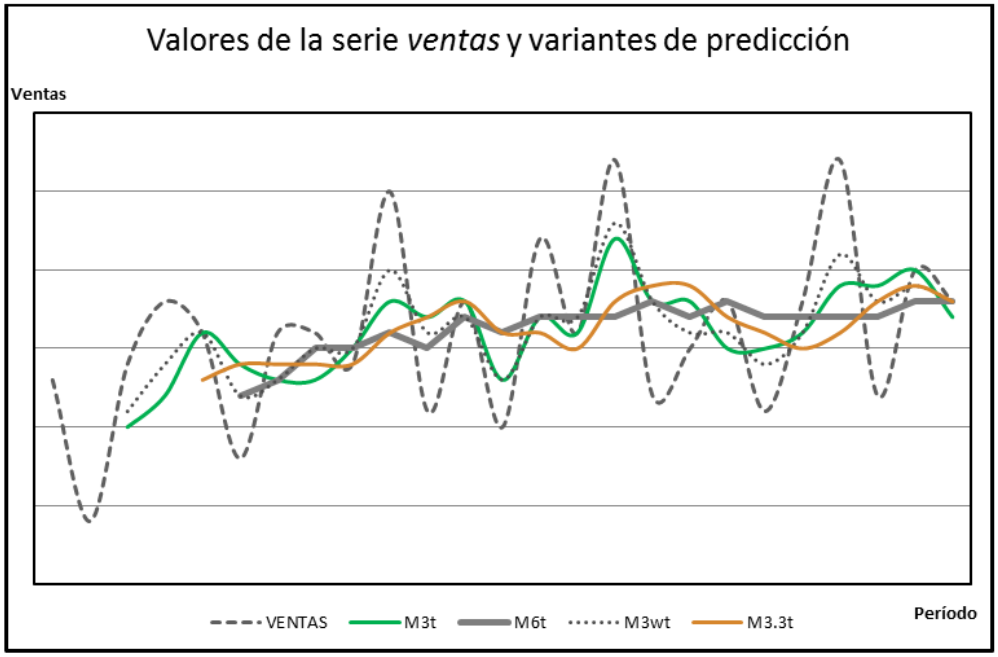


Figura 5 Valores observados y pronosticados de la variable Ventas

Caso 5. Alisados con tendencia

En la tabla se observa una serie-índice con datos sobre costes laborales unitarios en la industria manufacturera en España, obtenidos de la antigua Dirección General de Previsión y Coyuntura del Ministerio de Economía y Hacienda, la que se denomina COSLA. Con esta serie se debe hacer lo siguiente:

1. Importarla en EViews.

2. Analizar las características de la serie, en especial tendencia y estacionalidad.

3. Aplicar a la serie el alisado exponencial doble de Brown y Holt-Winters, con los parámetros óptimos. Comprobar con cuál se obtiene un menor error cuadrático medio. Comparar la predicción de ambos.

4. Aplicar el alisado exponencial simple y comparar los resultados con los obtenidos por las dos versiones del alisado doble. Aplicar el alisado exponencial doble con cualquier otro valor de alpha.

IMPORTACIÓN DE DATOS.

Para preparar la importación de datos, lo primero será crear un fichero de trabajo. Optamos por definirlo para datos sin referencia temporal específica de fecha, es decir, ordenados de 1 en adelante y para un máximo de 20 (17 observaciones muestrales y 3 datos de predicción, correspondiendo, realmente, el primer dato al año 1981 y el último a 1997).

Para ello, en EViews 7.1, seleccionamos en el menú principal: FILE / NEW / UNSTRUCTURED/ UNDATED, y aquí indicamos que trabajaremos con 20 datos. En WORFILE NAMES se indica el nombre de la serie COSLA

	COSLA
1	100,0000
2	110,5739
3	118,1201
4	126,3980
5	131,8828
6	137,6763
7	144,2019
8	149,4524
9	159,3335
10	176,4203
11	189,4843
12	205,1086
13	215,6812
14	210,4928
15	208,6748
16	219,5314
17	222,6703

Figura 6 Índice de costos laborales

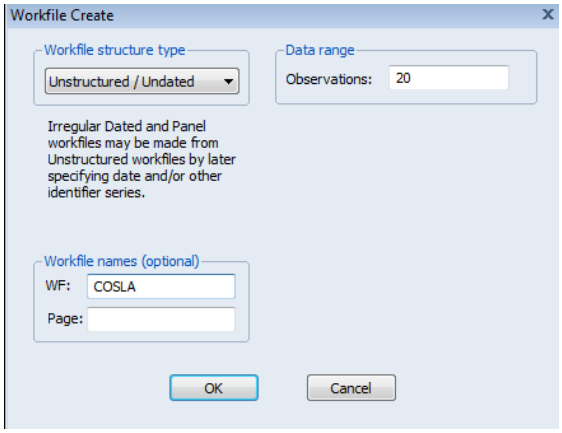


Figura 7 Generar archivo de trabajo en Eviews

B) CARACTERÍSTICAS DE LA SERIE

En el Workfile se selecciona la serie COSLA y con la opción VIEW / GRAPH / LINE & SIMBOL, se visualiza un gráfico de la serie, donde se comprueba la acusada tendencia que presenta.

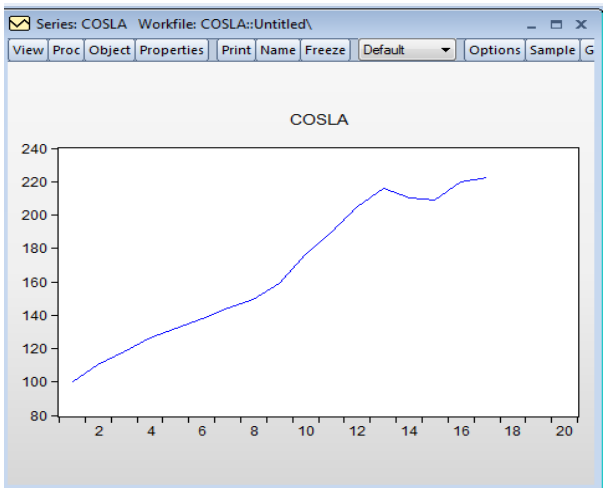


Figura 8 Evolución del índice de costos laborales

C) ALISADO EXPONENCIAL DOBLE DE BROWN Y HOLT-WINTERS.

Se procede a la estimación del alisado exponencial doble de Brown y de Holt-Winters, recogiendo los resultados, incluido predicciones de las nuevas series generadas en el proceso y que se han denominado, respectivamente, PCOSLAD y PCOSLAN.

Tanto con doble alisado de Brown (PCOSLAD) como con Holt-Winters (PCOSLAN), los parámetros que minimizan el error cuadrático medio (ECM) toman un valor alto, incluso unitario en el segundo caso. La suma de cuadrados de los residuos y la raíz cuadrada del ECM dan valores más reducidos en el alisado de Holt-Winters. Se indica además, que el nivel de la serie en su último dato (periodo 17) es de 222, con un incremento respecto al dato anterior (tendencia en ese punto) de unos 6 puntos.

Los valores de predicción se ajustarían muy estrechamente a los datos en el periodo histórico y mantendrían una tendencia similar en ambas técnicas a efectos de predicción. Se puede representar gráficamente la serie original (COSLA) y las dos series obtenidas en cada alisado (PCOSLAD y PCOSLAN), abarcando tanto el periodo histórico como el de predicción.

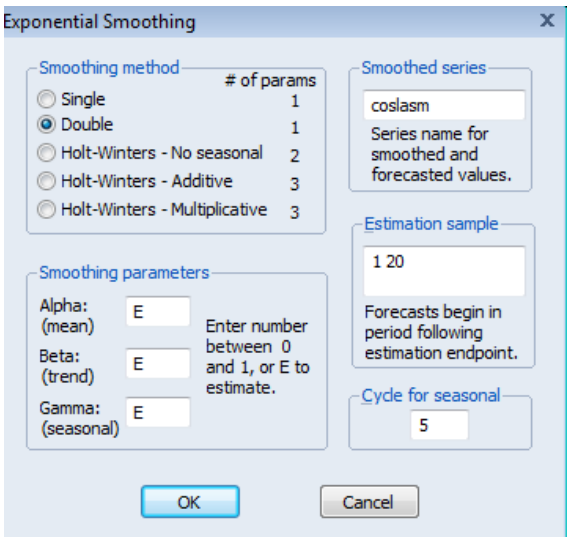


Figura 9 Configurar alisado exponencial en Eviews

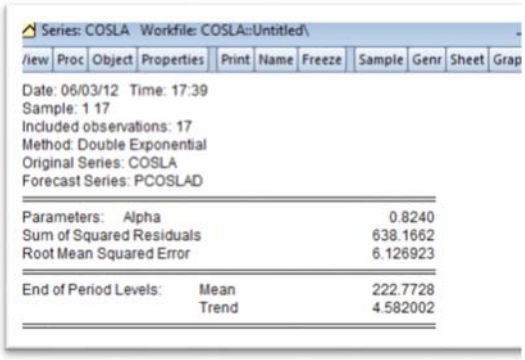


Figura 10

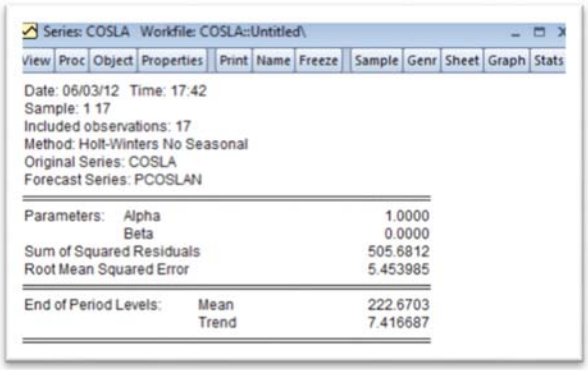


Figura 12

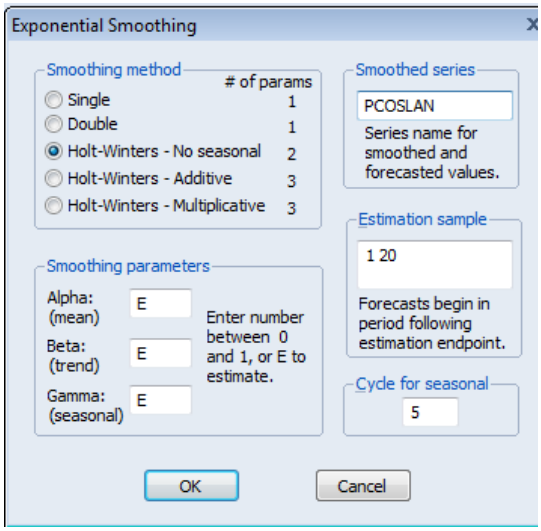


Figura 11

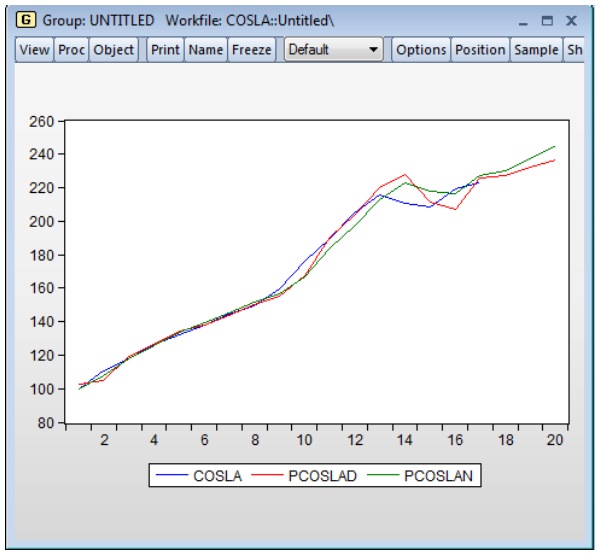


Figura 13

D) ALISADO EXPONENCIAL SIMPLE.

Naturalmente, los resultados son muy distintos a los que se obtendrían con un alisado exponencial simple, donde se comprueba cómo el error de estimación ha aumentado considerablemente.

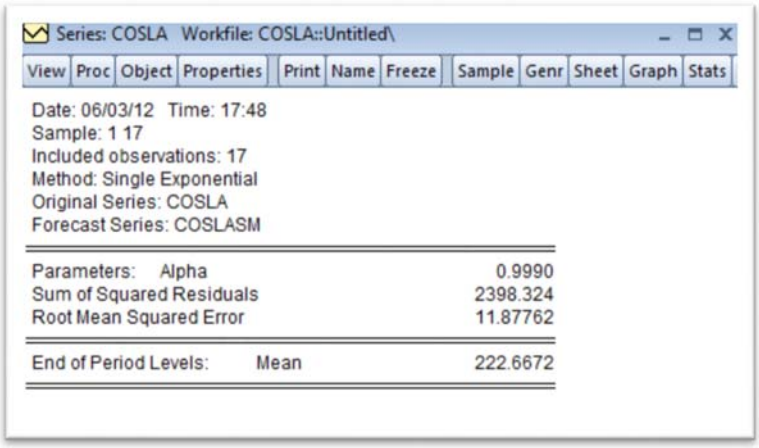


Figura 14

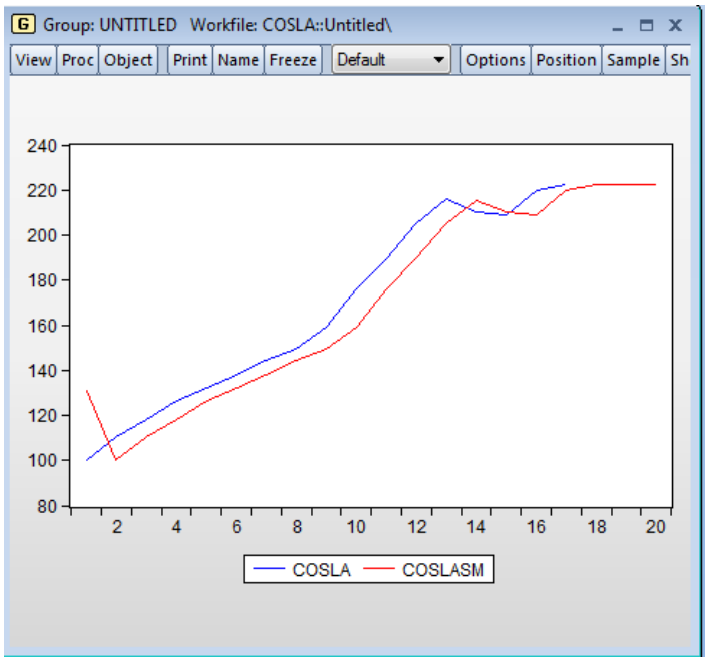


Figura 15

E) ALISADO EXPONENCIAL DOBLE CON VALORES NO ÓPTIMOS.

Por último, se puede probar con valores prefijados (no óptimos) de los parámetros, a fin de analizar la incidencia de estos cambios. Se aplica este análisis al alisado exponencial doble, con valores de alpha de 0,1, 0,5 y 0,9, generando las variables PCOSLAD1, PCOSLAD5 y PCOSLAD9.

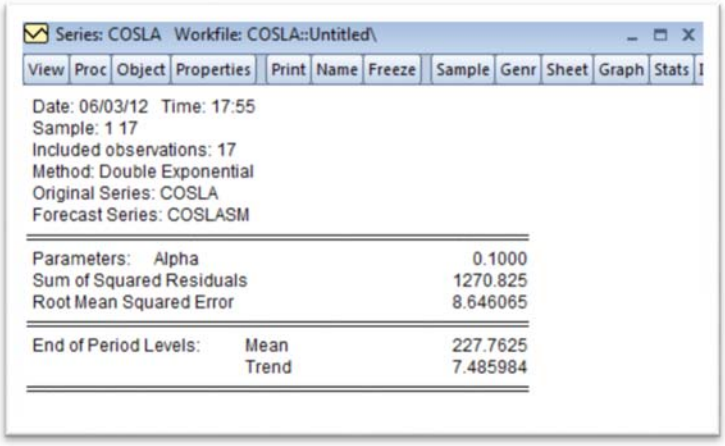


Figura 16

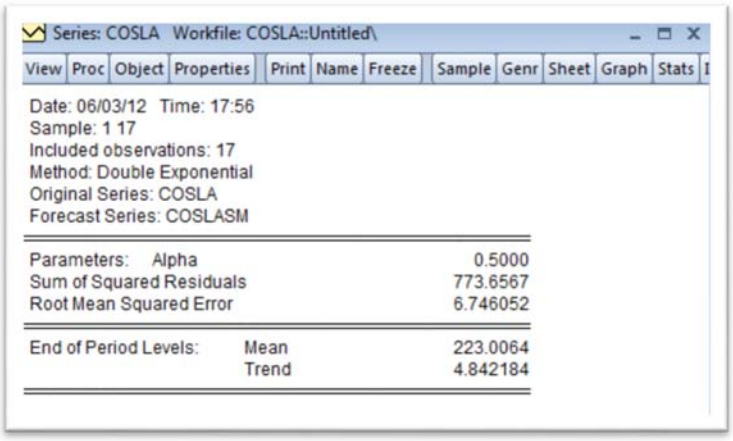


Figura 17

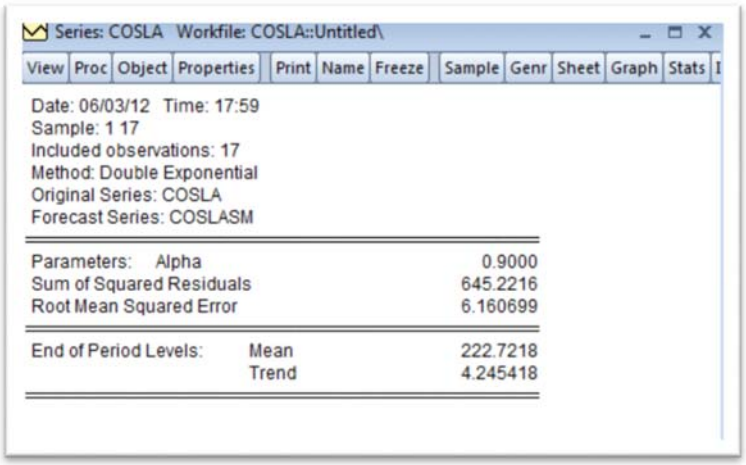


Figura 18

En la tabla siguiente se comparan los resultados obtenidos en cada alisado exponencial, en la que también incluimos el alisado exponencial de Holt-Winters, aparte de los datos originales y las predicciones supuestamente óptimas, en la que se aprecia cómo el error de estimación (ECM) aumenta con las series alisadas mediante un parámetro no óptimo.

	COSLA	PCOSLAD	COSLASM1	COSLASM5	COSLASM9	PCOSLAN
1	100	103,0213	103,0213	103,0213	103,0213	100
2	110,5739	104,999	109,3739	106,9569	104,5398	107,4167
3	118,1201	119,0919	116,5406	116,7754	119,9108	117,9906
4	126,398	126,181	123,7951	125,2259	126,0848	125,5368
5	131,8828	134,5694	131,2701	133,84	134,5953	133,8147
6	137,6763	138,32	138,3731	139,6178	137,9132	139,2995
7	144,2019	143,6132	145,2204	144,922	143,4901	145,093
8	149,4524	150,5003	151,9963	150,9622	150,5828	151,6186
9	159,3335	155,09	158,457	156,0327	154,9361	156,8691
10	176,4203	167,6885	165,5763	165,5363	168,3238	166,7502
11	189,4843	190,565	174,6978	183,4483	191,9318	183,837
12	205,1086	203,1992	184,7163	199,2333	203,1188	196,901
13	215,6812	220,0273	196,0039	216,3666	220,3105	212,5253
14	210,4928	227,8427	207,3523	228,408	227,1995	223,0979
15	208,6748	211,2768	215,5902	223,0483	208,5995	217,9095
16	219,5314	207,2353	221,8483	216,7515	206,6747	216,0915
17	222,6703	225,9793	228,9569	224,0147	227,8174	226,9481
18		227,3548	235,2485	227,8486	226,9672	230,087
19		231,9368	242,7344	232,6908	231,2126	237,5037
20		236,5188	250,2204	237,533	235,458	244,9204
ECM		638,1662	1270,825	773,6567	645,2216	505,6812

Figura 19

Caso 6. Análisis y descomposición de una serie temporal en EViews 7.1

Se tiene una serie temporal con datos sobre ventas del Estimador mensual industrial (EMI) de Argentina para 60 meses (desde Enero de 2005 a Diciembre de 2009, base 2004=100). A partir de los datos recogidos en la Figura 20, se pide:

Introduzca en EViews los datos de la serie. Analice con un gráfico la presencia clara de los componentes teóricos de una serie.

Obtenga el componente estacional a partir del método de la relación a la media móvil.

Compruebe con un gráfico la presencia de tendencia en la serie corregida de estacionalidad, y aíslela a partir de la aplicación de diferencias sucesivas.

Obtenga alternativamente la componente estacional con el método X-11, y compare los resultados.

A) INTRODUCCIÓN DE DATOS Y ANÁLISIS DE LOS COMPONENTES.

En primer lugar se define el workfile, con una frecuencia mensual de rango enero de 2005 a Febrero de 2008. Una vez definido el archivo de trabajo, se crea un objeto serie, y se introducen los datos, bien sea dato a dato (picado de datos) o importando la serie de Excel a través de la función explícita de EViews.

En el menú del workfile se puede seleccionar la opción SHOW para abrir la ventana del objeto serie, o bien se puede pulsar dos veces sobre la misma con el botón izquierdo del ratón. La apariencia que tiene la

ventana de la serie EMI es similar a la de una hoja de cálculo, con la información dispuesta en celdas (Figura 14.42). Se tiene ahora un nuevo menú específico de esta ventana. Si, por ejemplo, se elige la opción EDIT se puede editar la hoja de datos e introducir modificaciones en la misma.

Para visualizar un gráfico de la misma, se accede a la opción VIEW / GRAPH / LINE & SIMBOL que permite obtener un gráfico de evolución temporal de la serie en el período muestral indicado previamente.

La observación del gráfico de la serie EMI, en la Figura 14.43, permite apreciar la presencia de una componente de estacionalidad clara en la serie, identificada en las "disminuciones" que se presentan en los meses de Enero respecto al comportamiento medio de los meses restantes. Esto significa que en el primer mes del año se comprueba una menor actividad industrial. Por otra parte, la evolución general de la serie, es decir, a medio/largo plazo muestra una tendencia creciente. De hecho, puede imaginarse una línea recta con pendiente creciente en torno a la cual fluctúan los valores de la serie.

Para la aplicación de determinadas técnicas de predicción, en particular las denominadas elementales y en situaciones con historia (medias móviles y algunos tipos de alisados exponenciales) se debe proceder, previamente, a la transformación de los datos de la serie, para aislar la tendencia y estacionalidad.

B) ESTACIONALIDAD CON RELACIÓN A LA MEDIA MÓVIL.

Dado que la serie presenta tendencia y estacionalidad se procederá a la transformación de la misma. En primer lugar, la desestacionalización de la serie, procedimiento que consiste en la obtención de un factor de estacionalidad propio de cada mes. Para ello se supone un planteamiento multiplicativo de descomposición de la serie. De esta

forma, la serie original (EMI) será igual a una serie desestacionalizada (EMISA) multiplicada por un factor (FACTOR) (Figura 14.44):

EMI = EMISA * FACTOR

Mes	EMI	Mes	EMI
1	98,6	31	122,08
2	95,24	32	133,87
3	108,29	33	133,53
4	106,58	34	136,52
5	108,35	35	136,81
6	106,14	36	131,19
7	108,79	37	122,17
8	112,61	38	118,07
9	113,04	39	128,28
10	114,69	40	132,93
11	114,06	41	133,45
12	109,76	42	124,63
13	102,12	43	133,33
14	104,69	44	139,13
15	116,96	45	141,25
16	114,94	46	140,08
17	116,88	47	136,77
18	116,62	48	134,57
19	119,34	49	116,74
20	121,77	50	116,33
21	122,72	51	127,07
22	124,48	52	131,27
23	124,33	53	131,21
24	119,64	54	125,38
25	108,5	55	131,28

26	111,96	56	137,13
27	125,21	57	141,33
28	122,59	58	142,13
29	124,96	59	142,28
30	122,7	60	148,51

Figura 20 EMI Enero 2005 a diciembre 2009. Base 2004=100

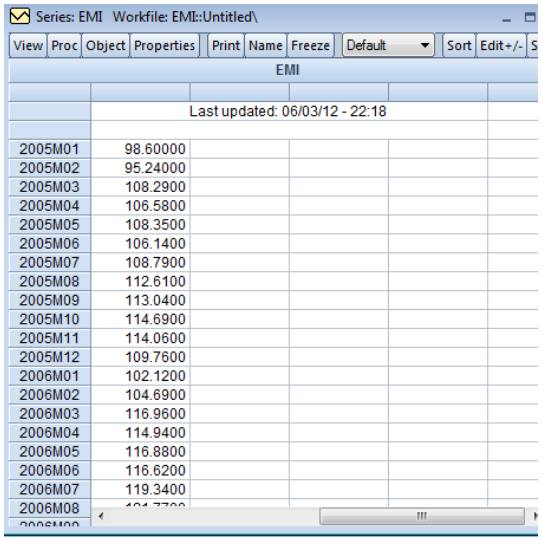


Figura 21

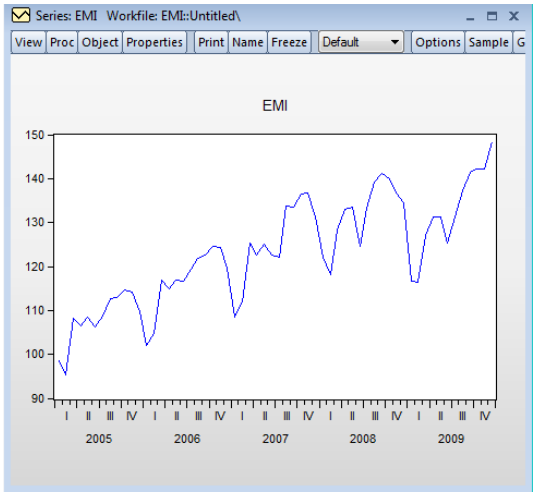


Figura 22

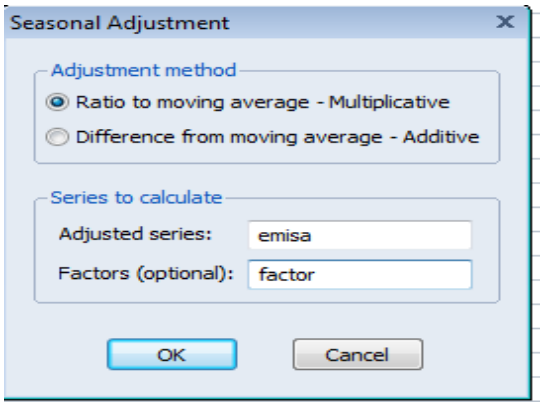


Figura 23

lo que significa que obtiene la serie desestacionalizada (EMISA) dividiendo los datos de la serie original entre un factor que contiene doce datos mensuales (uno diferente para cada mes).

Para realizar esta operación, en el menú principal de EViews, se abre la serie y se selecciona PROC / SEASONAL ADJUSMENT, y se elige el método de desestacionalización (Adjustment Method), indicando el nombre para la variable desestacionalizada (Adjusted Series) y para el factor (Factors).

Se selecciona el método, ya descrito en el texto, de relación a la media móvil (Ratio to moving average- Multiplicative). Se nombra a la serie desestacionalizada como EMISA, proporcionada por el Eviews por defecto. Al factor que calcula el programa se le llama, simplemente, FACTOR.

Como resultado se muestra la Figura 14.45 en la que se observan los datos del factor de estacionalidad calculado, donde se trata de valores en torno a la unidad, siendo más bajo el valor correspondiente al mes que presentaba inicialmente estacionalidad (Enero).

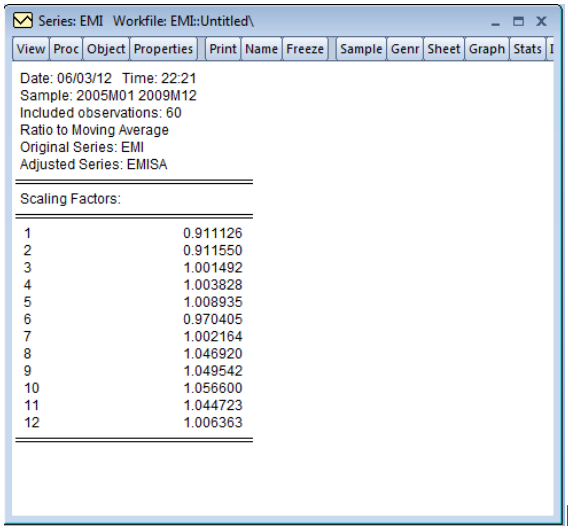


Figura 24

Dado que EViews ha generado la nueva serie desestacionalizada (EMISA) se puede visualizar un gráfico de la misma en la Figura 14.46, en forma similar a como se hizo con la serie EMI en la Figura 14.43.

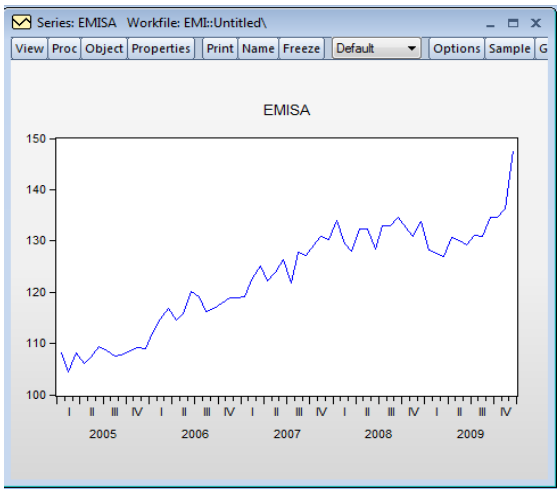


Figura 25

C) TENDENCIA CON DIFERENCIAS.

El gráfico de la serie EMISA muestra fluctuaciones que ya no se corresponden con la estacionalidad, puesto que ésta ha sido corregida y como es lógico sigue presentando tendencia. Una forma sencilla, y la más empleada, para corregir la tendencia consiste en la aplicación de primeras diferencias de la serie.

Para eliminar la tendencia a la serie EMISA, ya corregida de estacionalidad, se selecciona en el menú del workfile la opción GENR (Figura 14.47) y se escribe la operación a realizar, generando una nueva serie DEMISA a partir de las primeras diferencias de EMISA (Figura 14.48).

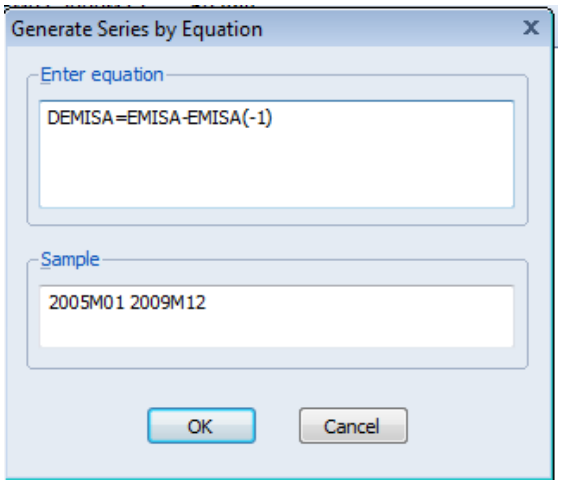


Figura 26

Se puede visualizar un gráfico, en la Figura 14.49, de la nueva serie DEMISA, serie sin estacionalidad y sin tendencia. Para comprobar que los datos de la serie transformada fluctúan en torno a un mismo valor, se calcula la media de esta serie generando (GENR) una variable denominada MEDIA, tal que:

MEDIA=@MEAN(DEMISA)

En el gráfico se representa la serie DEMISA y su valor medio (MEDIA). Para ello, en la opción SHOW en el menú de la ventana del workfile se indica ver las dos series mencionadas (Figura 14.48). EViews considera que se trata de un “grupo” de series y nos presenta ambas series en una ventana denominada GROUP a la que se le puede dar un nombre (NAME) si así se prefiere.

Para ver el gráfico con la representación de la serie DEMISA y su MEDIA, se procede de la misma forma que se hizo con los gráficos anteriores, pero ahora considerando las dos series. El gráfico sugiere una oscilación de los valores de la serie DEMISA en torno al valor medio, por lo que se considera ya como una serie sin tendencia. Es importante comentar aquí que el menú de la ventana de una serie y el menú de un grupo de series difieren, como es lógico, en su contenido.

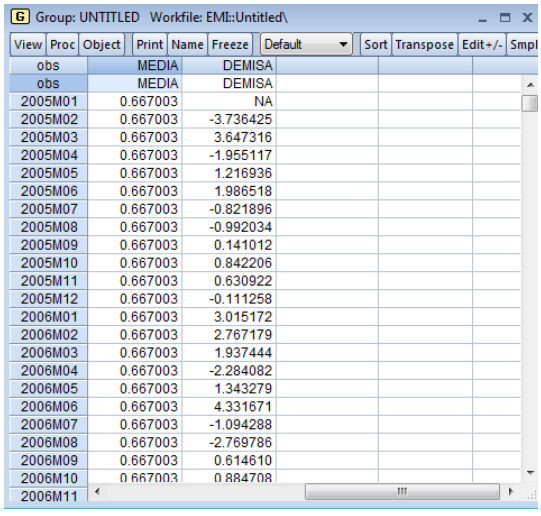
Las diferencias se encuentran, fundamentalmente, en el contenido de los submenúes VIEW y PROCS.

D) ESTACIONALIDAD CON X-11.

Finalmente, y a modo de comparación, se ha desestacionalizado la serie original, EMI, con otro procedimiento de los disponibles en EViews: el método CensusX11.

Como anteriormente, se selecciona en el menú principal PROC / SEASONAL ADJUSTMENT. La serie original a desestacionalizar, es EMISA, y la serie desestacionalizada se llamará EMISAX, siendo FACTORX el nombre elegido para el factor estacional con el método CensusX11-Multiplicativo (Figura 14.50).

Una vez obtenida la nueva serie, EMISAX, desestacionalizada con el procedimiento Census X11, se procede a representarla en un gráfico (Figura 14.51) junto con la serie EMISA, desestacionalizada con el método de la media móvil. Con la opción SHOW, del menú del workfile, se abre un grupo que contenga ambas series y en el menú de esta ventana de grupo se elige VIEW / GRAPH / LINE & SIMBOL



obs	MEDIA	DEMISA
obs	MEDIA	DEMISA
2005M01	0.667003	NA
2005M02	0.667003	-3.736425
2005M03	0.667003	3.647316
2005M04	0.667003	-1.955117
2005M05	0.667003	1.216936
2005M06	0.667003	1.986518
2005M07	0.667003	-0.821896
2005M08	0.667003	-0.992034
2005M09	0.667003	0.141012
2005M10	0.667003	0.842206
2005M11	0.667003	0.630922
2005M12	0.667003	-0.111258
2006M01	0.667003	3.015172
2006M02	0.667003	2.767179
2006M03	0.667003	1.937444
2006M04	0.667003	-2.284082
2006M05	0.667003	1.343279
2006M06	0.667003	4.331671
2006M07	0.667003	-1.094288
2006M08	0.667003	-2.769786
2006M09	0.667003	0.614610
2006M10	0.667003	0.884708
2006M11		

Figura 27

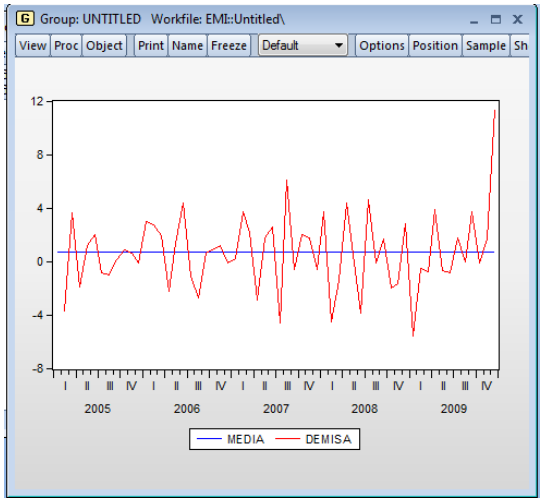


Figura 28

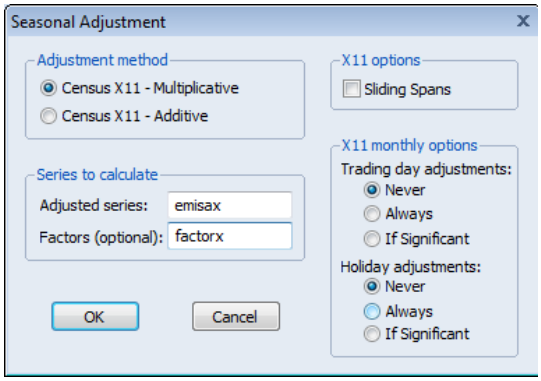


Figura 29

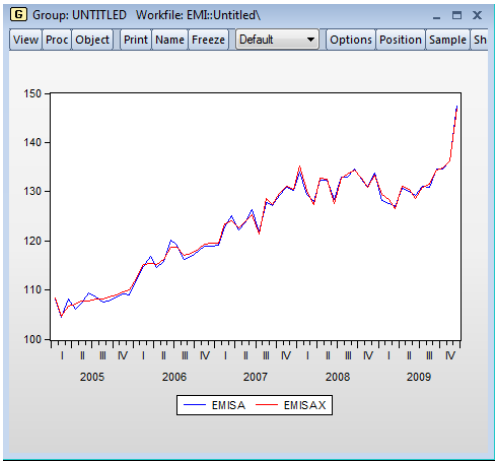


Figura 30

La evolución que presentan ambas series es similar, aunque parece que la serie EMISA suaviza más las oscilaciones de los valores de la producción industrial corregida de estacionalidad. Posiblemente sea más ilustrativo comprobar conjuntamente los valores de los dos factores de estacionalidad que han dado lugar a ambas series, donde:

EMISA = EMI / FACTOR

Mientras que,

EMISAX = EMI / FACTORX

Se compara los valores de los dos factores EN LA Figura 14.52.:

Date: 06/03/12 Time: 23:07
Sample: 2005M01 2009M12
Included observations: 60
Ratio to Moving Average
Original Series: EMI
Adjusted Series: EMISA

Scaling Factors:	
1	0.911126
2	0.911550
3	1.001492
4	1.003828
5	1.008935
6	0.970405
7	1.002164
8	1.046920
9	1.049542
10	1.056600
11	1.044723
12	1.006363

Figura 31

D10. FINAL SEASONAL FACTORS

YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
19**	90.927	90.932	101.573	99.537	100.528	98.430	100.664	104.077	104.198	105.147	104.157	99.843
19**	90.918	90.860	101.366	99.744	100.598	98.245	100.614	104.110	104.474	105.268	104.234	100.145
19**	90.730	90.629	100.903	100.016	100.676	97.918	100.546	104.115	104.802	105.341	104.284	100.490
19**	90.398	90.697	100.665	100.086	100.610	97.667	100.444	104.209	105.031	105.462	104.443	100.887
19**	90.134	90.623	100.443	100.099	100.556	97.474	100.360	104.265	105.143	105.480	104.460	101.043

Figura 32 FACTORX – Census X-11

La principal diferencia entre el factor calculado por el método de Relación a la media móvil y el del método CensusX11 es que, en este segundo caso, el factor puede variar año a año; mientras que, en el primer caso se asume un valor constante –observar que Eviews presenta al factor por el método a la media móvil como tasa y no como índice-. En líneas generales, ambos factores son muy parecidos, simplemente se aprecia un valor menor en los datos calculados para cada mes por el factor del método CensusX11 (FACTORX). Realmente, se podría utilizar indistintamente cualquiera de los dos procedimientos disponibles en EViews.

Como se ha señalado, el factor de estacionalidad del método de Relación a la media móvil (FACTOR) es el mismo en cada año. Sin embargo, si se visualiza los datos del factor de estacionalidad el método CensusX11 (FACTORX) para todo el periodo muestral se comprueba que no permanecen constantes.

Caso 7 Cálculo de tendencia mediante funciones matemáticas en EViews

Busque en el INDEC la serie del PBI trimestral de Argentina (base 1993=100) desde el primer trimestre de 1993 hasta el 4º trimestre de 2011.

Ajuste los datos a una tendencia a través de una recta.
Realice la predicción para el primer trimestre de 2012.
Ajuste los datos a una parábola de segundo grado, y prediga el valor del PBI para el primer trimestre del año 2012.

Caso 8 Predicción en una serie con componente estacional: el consumo de energía eléctrica.

El consumo de electricidad es un indicador parcial de actividad económica de gran relevancia en predicción macroeconómica y sectorial. En general, el consumo eléctrico mes a mes puede ser explicado en forma bastante operativa por el efecto conjunto de cuatro tipos de causas:

a) Estacionalidad propia del mes de que se trata.
b) Condiciones climatológicas excepcionales.
c) Variaciones en el número de días laborales o especiales.
d) Situación económica del país en su conjunto o de los principales sectores demandantes de electricidad.

En este caso de aplicación se trabaja con un indicador de consumo de energía obtenido a partir de la transformación de la serie original de Red Eléctrica de España para el periodo comprendido entre enero de 1981 y octubre de 1998 (Figura 14.58). Estas transformaciones realizadas pretenden tener presente aspectos como el número de días

de cada mes, efectos de laboralidad, temperatura o situación económica.

Utiliza el programa EViews y aplica las distintas técnicas de desestacionalización para la posterior predicción de los valores futuros de este indicador.

Año/Mes	Elec	Año/Mes	Elec	Año/Mes	Elec	Año/Mes	Elec	Año/Mes	Elec
1981M01	8794	1984M11	9363	1988M09	10065	1992M07	11535	1996M05	12255
1981M02	8880	1984M12	10039	1988M10	9809	1992M08	10257	1996M06	12933
1981M03	8577	1985M01	10053	1988M11	10973	1992M09	11416	1996M07	13119
1981M04	7859	1985M02	10207	1988M12	11987	1992M10	10964	1996M08	11796
1981M05	7614	1985M03	9288	1989M01	12305	1992M11	12071	1996M09	12542
1981M06	7679	1985M04	8882	1989M02	11965	1992M12	12661	1996M10	12542
1981M07	7744	1985M05	8545	1989M03	11180	1993M01	13059	1996M11	13690
1981M08	6239	1985M06	8751	1989M04	10636	1993M02	12680	1996M12	14658
1981M09	7524	1985M07	8882	1989M05	10117	1993M03	12062	1997M01	15027
1981M10	7703	1985M08	7621	1989M06	10386	1993M04	11184	1997M02	14366
1981M11	9313	1985M09	8905	1989M07	10633	1993M05	10782	1997M03	13866
1981M12	9604	1985M10	8836	1989M08	9259	1993M06	11224	1997M04	13279
1982M01	9654	1985M11	9786	1989M09	10369	1993M07	11491	1997M05	12937
1982M02	9265	1985M12	10578	1989M10	10337	1993M08	10252	1997M06	13212
1982M03	8588	1986M01	10876	1989M11	11572	1993M09	11240	1997M07	13486
1982M04	8123	1986M02	10612	1989M12	12495	1993M10	11283	1997M08	12426
1982M05	7681	1986M03	9585	1990M01	12747	1993M11	12334	1997M09	13619
1982M06	7815	1986M04	9244	1990M02	12485	1993M12	13193	1997M10	13349
1982M07	7894	1986M05	8729	1990M03	11595	1994M01	13447	1997M11	14430
1982M08	6370	1986M06	9033	1990M04	11008	1994M02	13180	1997M12	16020
1982M09	7829	1986M07	9173	1990M05	10519	1994M03	12333	1998M01	15996
1982M10	7672	1986M08	7665	1990M06	11076	1994M04	11841	1998M02	15645
1982M11	8693	1986M09	8907	1990M07	11051	1994M05	11361	1998M03	14559
1982M12	9421	1986M10	8738	1990M08	9587	1994M06	11893	1998M04	13956
1983M01	9530	1986M11	9854	1990M09	10883	1994M07	12083	1998M05	13430
1983M02	9100	1986M12	10481	1990M10	10507	1994M08	10920	1998M06	14111
1983M03	8787	1987M01	10862	1990M11	11813	1994M09	11915	1998M07	14659
1983M04	8352	1987M02	10615	1990M12	12830	1994M10	11606	1998M08	13196
1983M05	7990	1987M03	9874	1991M01	12946	1994M11	12880	1998M09	14224
1983M06	8174	1987M04	9536	1991M02	12634	1994M12	13674	1998M10	13870
1983M07	8323	1987M05	9114	1991M03	11756	1995M01	14527		
1983M08	6752	1987M06	9437	1991M04	11050	1995M02	13664		
1983M09	8044	1987M07	9600	1991M05	10826	1995M03	13117		
1983M10	8104	1987M08	8037	1991M06	11076	1995M04	12433		
1983M11	9318	1987M09	9421	1991M07	11252	1995M05	12018		
1983M12	9916	1987M10	9309	1991M08	9900	1995M06	12399		

1984M01	10400	1987M11	10474	1991M09	11331	1995M07	12516		
1984M02	10000	1987M12	11667	1991M10	10999	1995M08	11497		
1984M03	9181	1988M01	12039	1991M11	12169	1995M09	12366		
1984M04	8772	1988M02	11214	1991M12	12952	1995M10	12042		
1984M05	8338	1988M03	10702	1992M01	13066	1995M11	13181		
1984M06	8471	1988M04	10078	1992M02	12896	1995M12	14124		
1984M07	8611	1988M05	9494	1992M03	12173	1996M01	14786		
1984M08	7442	1988M06	9655	1992M04	11589	1996M02	13758		
1984M09	8560	1988M07	10069	1992M05	11002	1996M03	13348		
1984M10	8480	1988M08	8600	1992M06	10930	1996M04	12687		

Figura 30 Consumo de energía eléctrica

Caso 9: Función consumo

Con información suministrada por el Ministerio de Economía de la Nación, especifica un modelo de regresión lineal para estimar la función consumo en Argentina.

Caso 10: Razas de perros

A partir de la tabla de datos generada en el Cuaderno 4

a. Seleccione dos variables y pruebe la hipótesis de independencia.
b. Identifique grupos de razas de perros que presentan características semejantes. En particular se solicita analizar el primer plano factorial, identificar las coordenadas y contribuciones de las modalidades activas, seleccionar la mejor partición, identificar individuos característicos en cada grupo y caracterizar las clases por las modalidades y los individuos.

Caso 11. Perfil socioeconómico de lectores

Con la información suministrada en la base de datos provenientes del estudio para editar un nuevo diario regional, disponible en www.econometricos.com.ar, realice el AFCM y elabore el perfil socioeconómico de los potenciales lectores del nuevo diario.

Preguntas

En la Tabla de Burt de perfiles horizontales sólo es válida la lectura en forma horizontal ¿porqué?

Problemas

1. Replica el ejemplo desarrollado al final del capítulo utilizando algebra matricial.

2. El gerente de marketing de una cadena de autoservicio quiere determinar el efecto del espacio, en las estanterías, sobre las ventas de alimentos para animales domésticos. Se seleccionó una muestra aleatoria de 12 autoservicios, de igual tamaño, cuyos resultados se muestran en la tabla.

a) Grafique el diagrama de dispersión
b) Especifique el modelo y utilice el método de mínimos cuadrados para estimar los coeficientes de la regresión
c) Interprete el significado de la pendiente en este problema
d) Prediga las ventas semanales promedio, de alimentos para animales domésticos, para un autoservicio con 8 metros de estanterías para esos alimentos.
e) Calcule el error estándar de la estimación
f) Calcule el coeficiente de determinación e interprete su significado

g) Calcule el coeficiente de correlación
h) Encuentre una estimación de intervalo con el 90% de confianza en las ventas semanales promedio de un autoservicio que tiene 8 metros de estantería para alimentos de animales domésticos
i) Con un nivel de significación del 0.10, ¿hay relación lineal entre el espacio en la estantería y las ventas?

Relación entre ventas y espacio en estantería		
Autoservicio	Espacio X (metros)	Ventas semanales Y (miles de pesos)
1	5	1.6
2	5	2.2
3	5	1.4
4	10	1.9
5	10	2.4
6	10	2.6
7	15	2.3
8	15	2.7
9	15	2.8
10	20	2.6
11	20	2.9
12	20	3.1

3. En una fábrica de automóviles, un estadístico quiere desarrollar un modelo para predecir el tiempo de entrega (número de días entre la fecha del pedido y la fecha de entrega) de automóviles nuevos cuyo pedido contempla numeroso equipo opcional. El estadístico cree que hay una relación lineal entre el número de opciones pedidas y el tiempo de entrega. Se selecciona una muestra aleatoria de 16 automóviles que arroja los resultados de la tabla

a) Grafique el diagrama de dispersión
b) Especifique el modelo y utilice el método de mínimos

cuadrados para estimar los coeficientes de regresión

c) Interprete el significado de la pendiente en este problema
d) Si se ordena un automóvil que tenga 16 opciones, cuántos días cree usted que tardarían para la entrega?.
e) Calcule el error estándar de la estimación
f) Calcule el coeficiente de determinación e interprete su significado
g) Calcule el coeficiente de correlación
h) Encuentre una estimación de intervalo con el 95% de confianza para el tiempo medio de entrega de un automóvil equipado con 16 opciones
i) Con un nivel de significación del 0.05, ¿hay relación lineal entre el número de opciones y el tiempo de entrega?

Relación entre tiempo de entrega y opciones ordenadas

Automóvil	Número de opciones X	Tiempo de entrega Y
1	3	25
2	4	32
3	4	26
4	7	38
5	7	34
6	8	41
7	9	39
8	11	46
9	12	44
10	12	51
11	14	53
12	16	58
13	17	61
14	20	64
15	23	66
16	25	70

4. Un ingeniero agrónomo quiere determinar el efecto de aplicar un fertilizante orgánico natural sobre el rendimiento de tomates. Se van a utilizar cinco cantidades diferentes de fertilizantes en 10 lotes equivalentes. Los niveles de fertilizante se asignan en forma aleatoria a los lotes arrojando los resultados de la tabla.

a) Grafique el diagrama de dispersión
b) Especifique el modelo y utilice el método de mínimos cuadrados para estimar los coeficientes de regresión
c) Interprete el significado de la pendiente en este problema
d) Prediga el rendimiento de tomates para un lote al cual se le han aplicado 15 kilos por hectárea de fertilizante orgánico natural.
e) Calcule el error estándar de la estimación
f) Calcule el coeficiente de determinación e interprete su significado
g) Calcule el coeficiente de correlación
h) Encuentre una estimación de intervalo con el 90% de confianza para el rendimiento promedio de tomates fertilizados con 15 kilos por hectárea de fertilizante orgánico natural
i) Con un nivel de significación del 0.05, ¿hay relación lineal entre la cantidad de fertilizante utilizado y el rendimiento de tomates?

Relación entre rendimiento y nivel de fertilizante aplicado

Lote	Cantidad de fertilizante X (kilos/hectárea)	Rendimiento Y (kilos)
1	0	6
2	0	8
3	10	11
4	10	14
5	20	18
6	20	23
7	30	25
8	30	28
9	40	30
10	40	34

5. En el periodo Marzo 1993 y Diciembre 2003, la varianza observada en el consumo de energía eléctrica alcanzó 815151.19 Gwh y la varianza en el PBI 374.07 miles de millones de pesos a valores de 1993. La covarianza entre ambas variables es de 6204.11. ¿Cuál es la correlación entre ambas variables?

6. El consumo de gas observado en la Argentina en el periodo Marzo de 1993 y Diciembre de 2003 registró una variación de 510030.89 millones de metros cúbicos y la covarianza con el consumo de energía eléctrica alcanzó, en el mismo periodo 566380.53. Con el dato de varianza en el consumo de energía eléctrica de 815151.19 Gwh, calcule el coeficiente de correlación entre el consumo de Gas y el consumo de energía eléctrica.

7. En la tabla de contingencia se muestran las 200 personas que entraron en una tienda de equipos de sonido de acuerdo con el sexo y la edad. Pruebe si las dos variables categóricas son independientes.

Clientes de la tienda de aparatos de sonido			
	Sexo		
Edad	Hombre	Mujer	Total
Menor de 30	60	50	110
30 y más	80	10	90
Total	140	60	200

8. En la tabla se muestran las reacciones de los votantes ante un nuevo plan de impuestos sobre bienes raíces, de acuerdo con su afiliación política partidaria. Con estos datos, construya una tabla de frecuencias esperadas suponiendo que no existe relación ante el plan fiscal.

Reacciones de los votantes ante un nuevo plan de impuestos				
	Reacción			
Afiliación partidista	A favor	Neutral	Se opone	Total
Partido A	120	20	20	160
Partido B	50	30	60	140
Otro	50	10	40	100
Total	220	60	120	400

9. En la tabla siguiente se presenta la reacción de los estudiantes ante la ampliación de un programa deportivo colegial, de acuerdo con la clase a la que pertenecen. División inferior indica que se trata de un

alumno de primer o segundo año y División superior señala que los alumnos se encuentran en el tercer o cuarto año, siendo ésta la división de clase. Pruebe la hipótesis nula de que la posición de clase y la reacción ante el programa deportivo son variables independientes, utilizando el nivel de significación del 5%.

Reacción de los estudiantes ante el plan deportivo de acuerdo a su generación			
	Generación		
Reacción	División inferior	División superior	Total
A favor	20	19	39
En contra	10	16	26
Total	30	35	65

10 Un nutricionista dividió aleatoriamente a 15 ciclistas en tres grupos de 5 cada uno. Los integrantes del primer grupo recibieron suplementos vitamínicos que añadieron a sus corridas durante las tres semanas siguientes. A los del segundo grupo se les indicó que, durante esas tres semanas, tomaran un tipo especial de cereal de grano entero rico en fibra. A los miembros del tercer grupo se les dijo que comieran como hacían normalmente. Transcurrido el periodo indicado, el nutricionista hizo que cada ciclista recorriera una distancia de 6 kilómetros en la que registraron los siguientes tiempos:

Grupo	Tiempos en el recorrido de 6 kilómetros				
	1	2	3	4	5
De las vitaminas	15,6	16,4	17,2	15,5	16,3
Del cereal rico en fibra	17,1	16,3	15,8	16,4	16,0
De control	15,9	17,2	16,4	15,4	16,8

Estos datos ¿son consistentes con la hipótesis de que ni las vitaminas ni el cereal rico en fibras afectan a la velocidad de los ciclistas? Utiliza un nivel de significación del 0,05.

Tabla de Contenido

Referencias

- Berenson, M., & Levine, D. M. (1993). *Estadística para Administración y Economía. Conceptos y Aplicaciones.* Mexico: McGraw-Hill/ Intermericano de México S.A.
- Berenson, M., & Levine, D. M. (1996). *Estadística Básica en Administración.* México: Prentice Hall.
- Box, G. E., Hunter, J. S., & Hunter, W. G. (2008). *Estadística para Investigadores. Diseño, innovación y descubrimiento (Segunda edición).* Barcelona: Editorial Reverté, S.A.
- Crivisqui, E. M. (1993). *Análisis Factorial de Correspondencias. Un instrumento de investigación en Ciencias Sociales.* Asunción: Centro de Publicaciones Universidad Católica Nuestra Señora de Asunción.
- Crivisqui, E. M. (2001-2002). *Iniciación a los métodos estadísticos exploratorios multivariados.* Bruxelles, Belgique: Laboratorio de Metodología de Tratamiento de datos. Université Libre de Bruxelles.
- Dagum, C., & Bee de Dagum, E. M. (1971). *Introducción a la econometría (10 ed.).* México: Editorial Siglo XXI.
- Daniel, W. W. (1999). *Bioestadística, base para el análisis de las ciencias de la salud. (Tercera ed.).* México: Editorial Limussa.
- Droesbeke, J.-J., & Fine, J. (1997). *Metodología de la Encuesta.* Concepción, Chile: L. d. M. d. T. d. D. Université Libre de Bruxelles, Belgique, Programme de Recherche et d´enseignement en Statistique Appliquée. Chile: Universidad de Concepción. .
- Escofier, B., & Pagès, J. (1992). *Análisis Factoriales Simples y Múltiples: Objetivos, Métodos e Interpretación.* Bilbao: Servicio Editorial de la Universidad del País Vasco.
- Gujarati, D. N. (2004). *Econometría (cuarta ed.).* México: McGraw-Gill Interamericana editores, S.A. de C.V.
- Gujarati, D., & Porter, D. (2010). *Econometría (Quinta Edición).* México: McGraw Hill.
- Hildebrand, D. K., & Ott, L. (1997). *Estadística Aplicada a la Administración y a la Economía.* Wilmington Delaware, E.U.A.: Addison-Wesley Iberoamericana, S.A.
- Kazmier, L., & Díaz Mata, A. (1993). *Estaística aplicada a la Administración y a la Economía.* México: Mc. Graw Hill.

- Kinnear, T., & Taylor, J. (1993). *Investigación de Mercado.* Mc. Graw Hill.
- Maddala, G. S. (1996). *Introducción a la Econometría (Segunda Edición).* Naulcapan de Juarez, México: Prentice Hall.
- Pulido San Román, A. (2004). *Curso combinado de Predicción y Simulación.*
- Pulido San Román, A., & López García, A. M. (1999). *Predicción y simulación aplicada a la economía y gestión de empresas.* Madrid: Pirámide.
- Ross, S. M. (2007). *Introducción a la Estadística.* Barcelona: Editorial Reverté.
- Scheaffer, R. L., Mendenhall, W., & R.Lyman, O. (2007). *Elementos de muestreo (Sexta ed.).* Madrid: Editorial Thomson.
- Wackerly, D., Mendenhall, W., & Scheaffer, R. (2008). *Estadística Matemática con Aplicaciones (Séptima edición).* México: Cencage Learning.