

# El Paquete Estadístico SPAD

por

**XAVIER MILLET**

Centre d'Informàtica  
Universitat de Barcelona  
Diagonal 645, 08028 Barcelona

## DESCRIPCION GENERAL

Spad (Sisteme Portable pur l'Analyses des Données) es un paquete estadístico general. Implementa métodos descriptivos y multidimensionales aplicables a grandes matrices de datos cuantitativos, cualitativos o textuales.

El programa incluye utilidades propias de un sistema de software general tales como: selección de variables y casos, recodificación de datos, utilización de ficheros de trabajo y auxiliares, uso de etiquetas para la identificación de las variables, tabulación y representaciones gráficas de los resultados. Pero sin lugar a dudas lo más interesante son una serie de análisis exploratorios multidimensionales basados en las técnicas desarrolladas a partir de los trabajos de Benzecri (1973, 1980, 1982). Las dos características más destacables de este Software son la potencia y especificidad de métodos en el tratamiento de información textual junto con la concordancia de los métodos multidimensionales reproducidos con trabajos de importante base teórica.

De forma general los métodos que presenta el paquete se pueden catalogar en los siguientes grandes grupos:

Métodos estadísticos descriptivos básicos.

Métodos factoriales.

Análisis de correspondencias simples y múltiples.

Análisis en componentes principales.

Análisis lineal discriminante.

Métodos de clasificación.

El empleo de varios tipos de técnicas combinadas (factorización y clasificación) de forma muy fácil es lo que da al paquete unas grandes posibilidades de análisis.

## **ESTRUCTURA DEL PAQUETE**

El paquete está organizado de forma modular, a partir de las denominadas «ETAPES» o conjunto de instrucciones que engloban una función tal como: lectura de datos, factorización de matrices, almacenado de resultados intermedios, clasificaciones, representaciones...

La combinación de estas «ETAPES» es lo que da gran flexibilidad al paquete y permite crear ficheros de ejecución para los diferentes tipos de análisis, que abarcarán desde la descripción de una variable cualitativa hasta el tratamiento de datos textuales.

### **Un ejemplo: Descripción de una variable Cualitativa**

En este caso el análisis persigue la caracterización en clases de individuos diferenciados por una variable cualitativa. De forma complementaria el estudio se puede completar con el análisis de las correspondencias de la tabla de contingencias resultante de las clases de individuos generados y otras variables cuantitativas. Las etapas establecidas serán:

etapas:

DONNE	Lectura de los datos numéricos
LILEX	Selección de variables a utilizar.
TAMIS	Creación de las tablas con las variables cuantitativas a describir.
APLUM	Análisis de la tabla creada al cruzar la variable del paso anterior con otras variables cuantitativas.

## **ENUMERACION GENERAL DE LOS PRINCIPALES METODOS Y SUS «ETAPES»**

### **Tratamiento de datos tipo texto**

En el tratamiento de los datos textuales las funciones más generales con sus distintas etapas serán:

Entrada	
Etapas	
AFLEX	Del texto

---

DOTEX	De respuestas
DISCO	Creación de ficheros de trabajo
CORTE	Gestión y rectificación de las «palabras»

#### Construcción de tablas de contingencias para analizar

##### Etapas

COLEX	Cálculo de frecuencias de los nombres en los textos
COLEX	Cálculo de frecuencias de las respuestas en los textos
TALEX	Cálculo de frecuencias de las respuestas y las variables

##### Análisis

##### Etapas

APLUM	De las tablas de contingencia de los nombres y los textos
ASPAR	De las tablas de contingencia de los nombres y las respuestas
APLUM	De las tablas de contingencia de los nombres y las variables

#### Caracterización de los textos

##### Etapas

MOCAR	Por nombres y respuestas típicas
ROCAR	Por respuestas típicas y distancias

#### Gráficas factoriales

##### Etapas

POLEX	Situación de la palabra en el espacio factorial
POSIT	Situación de las variables ilustrativas

#### Métodos y etapas generales

##### Entrada de datos

##### Etapas

DONNE	Entrada de datos numéricos
DPLUM	Entrada de datos en frecuencias
LILEX	Creación de un diccionario de variables y datos
ECRIT	Introducción de transformaciones de un diccionario

#### Gestión y utilidades de SPAD

##### Etapas

ECRIT	Gestión de un diccionario
LISTP	Listado de parámetros e instrucciones
CODAJ	Preparación de ficheros

ARCHI	Almacenado de resultados
TRANS	Creación de ficheros de trabajo (NLEG) a partir de (NDIC, NDON)
VALDA	Modificación de ausencias ("missing").

### Rectificación y Transformaciones

#### Etapas

CODAJ	Preparación de ficheros
CREAT	Rectificaciones generales
CROIS	Combinaciones de variables de grupos diferentes
FREGA	Codificación en clases de variables continuas
REDRE	Redefinición de una muestra
COMRO	Recodificación en rango estimas de rangos.

### Selección de datos

#### Etapas

LILEX	Para un tratamiento
ARCHI	Para guardar resultados

### Estadísticos básicos

#### Etapas

TRIH	Histogramas
TABUL	Tabulaciones
COMPL	Medias, desviaciones y correlaciones
MULTCT	Tablas de correspondencias
REPLA	Gráficas para dos coordenadas
DIS2G	Análisis discriminante en dos grupos
MCREG	Regresión y análisis de la varianza
FUSIL	Mejor regresión y análisis discriminante

### Clasificaciones

#### Etapas

ECLAT	Mixtas para datos nominales
SEMIS	Mixtas para coordenadas factoriales
RECIP	Jerárquicas para coordenadas factoriales
TAMIS	Caracterización de clases
INERC	Calidad de una partición
ARCHI	Para almacenar una partición

### Análisis factorial

#### Etapas

COMPL	En componentes principales
-------	----------------------------

CORBI	Correspondencias simples
APLUM	Correspondencias simples
MULTC	Correspondencias múltiples
MULDI	Correspondencias múltiples
CLAIR	Descripción detallada de un factor
ARCHI	Para almacenado de un factor

### Gráficas factoriales

#### Etapas

FPLAN	Asociadas a un análisis
GRAFK	Asociadas a una partición
POSIT	Para un grupo de variables nominales

## CONSIDERACIONES GENERALES

El paquete está escrito en Fortran dándole la característica de absoluta portabilidad. El tamaño total unas 40.000 líneas de programa fuente genera ficheros ejecutables grandes y poco optimizados, siendo la ejecución relativamente lenta. Es remarcable que el diseño del paquete es para el tratamiento y análisis de matrices de datos de grandes dimensiones. Ello significa que permite tratar tablas de 70 por 1600 de forma sencilla, y por tanto muy en concordancia con el uso habitual del tipo de métodos que implementa.

La sintaxis de las «Procedures» es sencilla pero rígida ya que sigue normas de posición fijas. Por todo ello el uso inicial del paquete no es muy práctico y se hace recomendable utilizar los ficheros de instrucciones ya creados limitándose a sustituir las instrucciones. La documentación disponible, ver apartado de bibliografía, es un tanto farragosa al mezclar métodos y el empleo del paquete. Se hecha muy en falta manuales de introducción o iniciación. Por el contrario, la bibliografía que describe las metodologías implementadas es de gran calidad.

La parte gráfica tanto para lo que hace referencia a los gráficos factoriales como a los de clasificación es muy útil y potente, y se combina muy adecuadamente con las utilidades básicas del paquete (identificadores y/o etiquetas...). Toda esta potencialidad gráfica es limitada a la impresión de resultados echándose en falta salidas de representación gráfica de calidad. No dispone de ninguna utilidad de intercambio directo con otros paquetes.

Los métodos estadísticos más comunes y presentes en la mayoría de paquetes se presentan en éste de forma rigurosa. Sin duda la característica diferencial del paquete es la incorporación de métodos no presentes en otros sistemas de software estadístico, destacando los métodos de correspondencia y el análisis y tratamiento de informaciones textuales.

La versión de SPAD denominada SPAD.N supera en algunas cuestiones a su predecesora flexibilizando la sintaxis de las instrucciones de comunicación con el paquete SAS. Es pero una versión parcial al no incluir la parte más novedosa y peculiar del paquete como es el tratamiento de datos textuales y encuestas.

## BIBLIOGRAFIA

- J. P. BENZECRI (1973). "*Analyse des Données*" Tome I; "*La Taxinomie*" Tome II; *Correspondances* Dunod - Paris.
- J. P. BENZECRI ET COLL. (1980). "*Linguistique et Lexicologie, Pratique de l'Analyse des Données*" Tome III - Dunod - Paris.
- J. P. BENZECRI (1982). "*Construction d'une Classification Ascendante Hiérarchique par la Recherche en Chaîne des Voisins Réciproques*" - *Cahiers de l'Analyse des Données*, Vol. VII, No. 2, pp. 209-218.
- J. M. BOUROCHE, G. SAPORTA (1983). "*L'Analyse des Données*" - P.U.F., Collection "*Que sais-je*" - Paris.
- F. CAILLEZ, J. P. PAGES (1976). "*Introduction à l'Analyse des Données*" - SMASH, 9, rue Duban 75016 Paris.
- E. DIDAY (1983). "*Eléments d'Analyse des Données*" - Dunod - Paris.
- M. GREENACRE (1984). "*Theory and Application of Correspondence Analysis*" - Academic Press - London.
- M. JAMBU, M. O. LEBEAUX (1978). "*Classification Automatique pour l'Analyse des Données*" Tome 1 - "*Méthodes et algorithmes*" Tome 2 - "*Logiciels*" Dunod - Paris.
- J. JUAN (1982). "*Programme de Classification Hiérarchique par l'Algorithme de la Recherche en Chaîne des Voisins Réciproques*" *Cahiers de l'Analyse des Données*, Vol. 7, No. 2, pp. 219-225.
- L. LEBART, A. MORINEAU, N. TABARD (1977). "*Techniques de la Description Statistique, Méthodes et Logiciels pour l'Analyse des Grands Tableaux*" - Dunod - Paris.
- L. LEBART, A. MORINEAU, J. P. FENELON (1979). "*Traitement des Données Statistiques, Méthodes et Programmes*" - Dunod - Paris.
- L. LEBART, A. MORINEAU, K. W. WARWICK (1984). "*Multivariate Descriptive Statistical Analysis, Correspondence Analysis and Related Techniques for Large Matrices*" - Wiley - New York.
- I. C. LERMAN (1981). "*Classification et Analyse Ordinale des Données*" - Dunod - Paris.
- L. L. MAC QUITTY (1966). "*Similarity Analysis by Reciprocal Pairs of Discrete and Continuous Data*" - *Educ. Psych. Measurement*, Vol. 26, pp. 825-831.
- A. MORINEAU (1982). "*Choice of Methods and Algorithms for Statistical Treatment of Large Arrays of Data*" - *Proceedings in Computational Statistics (COMPSTAT)*, Physica Verlag, Vienne.
- L. LEBART (1981). "*Une procédure d'Analyse Lexicale Ecrite en Langage FORTRAN*" - *Les Cahiers de l'Analyse des Données*, Vol. 6, No. 2, pp. 229-243.
- L. LEBART (1982). "*Exploratory Analysis of Large Sparse Matrices with Application to Textual Data*" - *COMPSTAT* - pp. 67-76.

- 
- L. LEBART (1982). "*L'Analyse Statistique des Réponses Libres dans les Enquetes Socio-Economiques*" - Consommation - No. 1, pp. 39-42 - Dunod - Paris.
- A. MORINEAU (1984). "*Computational and Statistical Methods for Exploratory Analysis of Textual Data*" - COMPSTAT - Physica-Verlag - Vienna.
- T. ALUJA, L. LEBART (1984). "*Local and Partial Principal Component Analysis*" - COMPSTAT - Physica-Verlag - Vienna.
- T. ALUJA, L. LEBART (1985). "*Factorial Analysis upon a graph*" - Bulletin Technique du CESIA - Vol. 3 - pp. 4-34. J. P. BENZECRI (1973) - "*L'Analyse des Données*" - Tome 2 - Dunod - Paris.
- A. D. CLIFF, J. K. ORD (1981). "*Spatial processes, Models and Applications*" - Pion limited - London.
- R. C. GEARY (1954). "*The contiguity ratio and Statistical mapping*" - The Inc. Statistician - p. 115-145.
- L. LEBART (1969). "*Analyse Statistique de la Contiguïté*" - Publ. de l'ISUP - XVIII - p. 81-112.
- L. LEBART, N. TABARD (1973). "*Recherches sur la Description Automatique des Données Socio-économiques*" - Report CREDOC-CORDES - No. 13/1971 - Chapitre III.
- L. LEBART (1978). "*Programme d'Agrégation sous Contraintes*" - Cahiers de l'Analyse des Données - Vol. 3, No. 3, pp. 275-287.
- L. LEBART (1984). "*Correspondence Analysis of Graph Structures*" - Bulletin technique du CESIA - Vol. 2, No. 1-2, p. 5-19.



# REVISTA ESTADISTICA ESPAÑOLA

## INFORMACION PARA LOS AUTORES

### Envío de originales

Los originales deben enviarse al Director de la Revista, INE, Paseo de la Castellana, 183. 28071 Madrid. Deben estar redactados en castellano y no haber sido publicados o estar en proceso de publicación en otro lugar. Se enviarán 4 copias del trabajo y la revista acusará siempre su recibo.

### PRESENTACION DE LOS MANUSCRITOS

#### Estructura

La primera página debe incluir exclusivamente el título del artículo y el nombre, dirección completa y teléfono del autor. En el caso de varios autores se indicará a quién debe dirigirse la correspondencia. La segunda página contendrá únicamente el título y resumen del trabajo de un máximo de 100 palabras, seguida de 3 a 6 palabras clave y la clasificación AMS del artículo. El texto del artículo comenzará en la tercera página y las secciones se numerarán consecutivamente. La última página del original contendrá en inglés el título del artículo, un resumen del mismo bajo el epígrafe Summary y las palabras clave. El manuscrito debe mecanografiarse a doble espacio. Cuando el título del artículo contenga más de 80 caracteres, se deberá indicar un título alternativo de dicha longitud o menor.

#### Gráficos

Todos los diagramas o gráficos se numerarán sucesivamente y se indicará su posición en el texto con el nombre de *figura*. Se colocarán al final del manuscrito y deberán ser de la calidad necesaria para su reproducción.

#### Referencia

El sistema de referencia a seguir es el oficial del International Statistical Institute. Los autores se citarán en el texto por su nombre, seguido de la fecha de publicación; ejemplo: Box (1986), y las referencias se situarán en orden alfabético al final del texto, como sigue:

- a) *Libros*: Weisberg, S. (1985). *Applied Linear Regression*, New York: Wiley.
- b) *Artículos*: Mahalanobis, P. c. (1950). Why Statistics? *Sankhya*, 10, 195-228.
- c) *Trabajos en obras colectivas*: Box, G. E. P. (1983). An Apology for Ecumenism in Statistics. *Scientific Inference, Data Analysis and Robutsness*, Ed. C. E. P., Leonard, T. y Wu, C. F., pp. 51-84. New York: Academic-Press.

#### Evaluación de los originales

Los originales serán sometidos a un proceso de evaluación garantizando el anonimato tanto del autor como de los evaluadores. El objetivo de la revista es que el autor reciba información sobre el resultado de la evaluación en un plazo máximo de tres meses.

#### Pruebas y separatas

Aceptado el artículo y antes de su publicación definitiva, el autor o autores recibirá 2 juegos de pruebas para correcciones, uno de los cuales deberá devolver corregido a la Revista en un plazo de una semana desde su recepción. Una vez publicado, recibirá 25 separatas de su trabajo. Los posibles costes de impresión derivados de cualquier modificación de la versión final aceptada del manuscrito o de retraso en la corrección de pruebas serán a cargo del autor o autores del mismo.