Anales de la Real Academia de Doctores de España
Volumen 12, pp. 39-72, 2008

ELEMENTOS BÁSICOS PARA UNA NUEVA ESTRATEGIA METODOLÓGICA EN EL TRATAMIENTO ESTADÍSTICO DE UNA ENCUESTA DE OPINIÓN

DR. FCO. JAVIER DÍAZ-LLANOS SAINZ-CALLEJA
Académico de Número de la RADE.
Medalla de Plata al Mérito Doctoral de la RADE.
Departamento de Medio Ambiente
Ministerio de Ciencia e Innovación(INIA)
Jefe Laboratorio de Estadística Ambiental.

DRA. MARÍA DEL CARMEN CERMEÑO CARRASCO
Ex-profesora e investigadora científica de Citogenética
de las Universidades de Munich y Libre de Berlín.
Referee de artículos en dichas Universidades.
Académica de Número de Genética Humana
de la Universidad de Navarra.

RESUMEN

Este trabajo trata de establecer una **nueva estrategia metodológica *ad hoc*** —con sus correspondientes etapas— para el **tratamiento estadístico de las encuestas de opinión,** de la manera más didáctica posible. Dicha **estrategia** está basada en el **compromiso** entre los **operadores WD de Yves Escoufier** y en las **técnicas actualizadas**, tanto de **métodos factoriales** como de **algoritmos de clasificación.**

Palabras clave

Tabla de datos lógica. Tabla de datos lógica desdoblada. Operadores WD de Yves Escoufier. Tablas de contingencia. T cuadrado de Tschuprow. Análisis de correspondencias simples (AFC). Análisis de correspondencias múltiples (ACM). Análisis en componentes principales (ACP). Análisis discriminante lineal (AFDL). Algoritmos de clasificación jerárquica (CAH) y no-jerárquica.

INTRODUCCIÓN

Existen —como es bien sabido— numerosos trabajos, libros, tesis doctorales y artículos sobre el **tratamiento estadístico de las encuestas de opinión.** Desde nuestro punto de vista resaltamos (1, 2, 3, 4, 5, 6, 7, 8), respectivamente.

Cabe destacar que la puesta a punto de una **nueva estrategia metodológica** basada en el **Análisis Estadístico Multidimensional Lineal, fuera de hipótesis distribucionales *a priori*** para el **tratamiento estadístico de las encuestas de opinión,** no es suficiente para llegar a unos resultados coherentes con la realidad, sino que hay que tener en cuenta otros componentes.

En este sentido, Pierre Dagnelie en un artículo —publicado en un periódico francés— titulado: **«Les pièges de la statistique»** hace hincapié de cuál debe ser el **proceso metodológico** para llegar con éxito a conclusiones satisfactorias acordes con la realidad.

Siguiendo los consejos de Pierre Dagnelie, contenidos en su artículo, proponemos tres puntos fundamentales —concatenados entre sí— que se aplicarían siguiendo la siguiente secuencia:

Primer punto

La **recogida de datos** debe realizarse con el **mayor rigor posible;** es decir, no sólo el cuestionario se diseñará —exclusivamente— por verdaderos especialistas en el tema concreto, objeto de estudio, sino además, las preguntas contenidas en el formulario deben ser entendidas —sin dificultad—, por aquel **segmento** de personas a las cuales vaya dirigido dicho cuestionario. Una vez cumplido este **punto,** estamos en condiciones adecuadas para el siguiente **punto.**

El **segundo punto** consiste en la **elección de una nueva estrategia metodológica,** basada en el **Análisis Estadístico Multidimensional Lineal, fuera de hipótesis distribucionales *a priori.***

En cuanto al **tercer** y último **punto,** refiere que el resultado de los análisis tengan un cierto grado —aceptable— de credibilidad.

Para conseguir esto es imprescindible que el estadístico trabaje conjuntamente —durante todo el **proceso metodológico**—, con **verdaderos especialistas que conozcan en profundidad los datos empíricos.**

En nuestro caso tan sólo haremos hincapié en el **segundo punto.**

Así pues, el objetivo de este artículo es el establecimiento de una **nueva estrategia metodológica *had hoc*** —con sus correspondientes etapas— para el **tratamiento estadístico de una encuesta de opinión,** de la manera más didáctica posible. Dicha **estrategia** está basada, tanto en los conceptos contenidos en (9, 10, 11) como en un conjunto de **métodos factoriales** y **algoritmos de clasificación actualizados, fuera de hipótesis distribucionales *a priori.***

Aquellas etapas no contenidas en los tratados clásicos de **Análisis Estadístico Multidimensional Lineal** —al menos— escritos en castellano, se explicarán de la manera más detallada posible. Los resultados obtenidos para **tablas de datos lógicas,** de dimensiones (n,p), se ilustrarán, asimismo, con la aplicación de dicha **estrategia** a una miniencuesta de opinión.

MATERIAL Y MÉTODO

Material

Encuestas de opinión

1. **Tabla de datos lógica de dimensiones (n,p).**
2. **Tabla de datos lógica de dimensiones (19, 4).**

La **nueva estrategia metodológica** que presentamos sólo es válida para este tipo de **tablas de datos.**

La **nueva estrategia metodológica** que presentamos sólo es válida para 1. Sin embargo, 2 nos es muy importante para la comprensión optimizada de sus **11 etapas,** así como la de su fase de **depuración** previa.

Pre-método

Antes de mostrar las **11 etapas** que constituyen la **estrategia metodológica** basada, no sólo en los conceptos contenidos en (9, 10, 11), sino también en un conjunto amplio de **métodos factoriales** y **algoritmos de clasificación, fuera de hipótesis distribucionales *a priori,*** tenemos que someter a la **tabla de datos lógica** y a la **tabla de datos,** conocida con el nombre de **tabla disyuntiva completa,** cuyas columnas vienen definidas por las **variables indicadoras** V11+, V12+, V21+, V22+,...,Vp1+, Vp2+, a un **proceso de depuración.**

Así pues, el número de **variables indicadoras** será 2p, en donde p representa el número de **variables cualitativas** a dos modalidades.

1. Proceso de depuración de una tabla de datos lógica

El proceso de depuración de una **tabla de datos lógica** está constituido por dos etapas.

Primera etapa

1.1. Eliminación de las filas y las columnas en las cuales todos sus valores son ceros

El hecho de la eliminación de dichas filas y columnas bajo la condición ya aludida es porque los cálculos que se realizan en un **AFC (método factorial constitutivo en la estrategia metodológica),** no permite esta situación.

1.2. Si las marginales de las filas de la tabla de datos lógica no son iguales, es aconsejable desdoblar cada una de las columnas de la tabla de datos lógica (12). Como resultado de este **desdoblamiento,** obtenemos una **tabla de datos** conocida con el nombre de **tabla disyuntiva completa.**

Segunda etapa

2. Proceso de depuración de una tabla disyuntiva completa

2.1. Eliminación de las columnas que no alcancen, al menos, un 2% de unos

El hecho de la eliminación de dichas columnas bajo la condición ya aludida es con el fin de conseguir la **estabilidad de todos los ejes factoriales** que se calculan en un **ACM** (6).

2.2. Eliminación de las variables cualitativas que no cumplan ciertas condiciones

Si como consecuencia de la eliminación de, al menos, una columna de la **tabla de datos disyuntiva completa,** al menos una de las **variables cualitativas** tiene sólo una **modalidad,** se procederá a la eliminación de dichas **variables cualitativas.**

MÉTODO

2. Estrategia metodológica

Una vez realizado el **proceso de depuración** en la **tabla de datos** que se desea **analizar** y, por consiguiente, haber retenido cuál es la **tabla de datos** para ser sometida con **técnicas de Análisis Estadístico Multidimensional Lineal, fuera de hipótesis distribucionales *a priori,*** mostraremos las etapas que constituyen la **nueva estrategia metodológica** que vamos a exponer para el **tratamiento estadístico de una encuesta de opinión,** con el propósito de **extraer** la **máxima información** de la **tabla de datos original.**

Primera etapa

La **primera etapa** consistirá en la aplicación de un **algoritmo de clasificación jerárquica descendente** a la **tabla de datos lógica** y a la **tabla de datos lógica desdoblada,** debidamente **depuradas** (véase **proceso de depuración de una tabla de datos lógica).** Como resultado de tal aplicación, podríamos observar las diferencias que exhiben cada una de las dos **tablas de datos** resultantes de la aplicación de dicho **algoritmo.**

Segunda etapa

La **segunda etapa** consistirá en la aplicación del **AFC** a la **tabla de datos lógica** y a la **tabla de datos desdoblada** debidamente **depuradas** (12).

Invitamos —muy vivamente— a los investigadores a que apliquen el **AFC** a sus propios **datos empíricos.** Estos podrán observar que existen ciertas diferencias entre los resultados obtenidos de la aplicación del **AFC** tanto a una como a otra **tabla de datos.**

Tercera etapa

La **tercera etapa** consistirá en la construcción de las [p(p-1)/2] **tablas de contingencia** obtenidas a partir de la **tabla disyuntiva completa,** cuyas columnas representan la 2p **variables indicatices:** V11+, V12+,...,Vp1+, Vp2+.

Cuarta etapa

La **cuarta etapa** consistirá en la construcción de las [p(p-1)/2] **T cuadrado de Tschuprow** a partir de las [p(p-1)/2] **tablas de contingencia** obtenidas en la **tercera etapa.**

Observación de interés

Las [p(p-1)/2] **T cuadrado de Tschuprow** son las mismas tanto para las **variables cualitativas** V1+, V2+,..., Vp+ como para las **variables cualitativas** V1–, V2–,..., Vp–, todas ellas a dos **modalidades.** Por tanto, retendremos tan sólo para el **análisis de datos** la **tabla disyuntiva completa,** proveniente de **variables cualitativas** V1+, V2+,..., Vp+, todas ellas a dos **modalidades.**

Esta **etapa** nos permitirá saber el **nivel de asociación** existente entre las **variables cualitativas** V1+, V2+,..., Vp+, a dos **modalidades.**

Quinta etapa

La **quinta etapa** consistirá en construir a partir de las [p(p-1)/2] **T cuadrado de Tschuprow,** las [p(p-1)/2] distancias entre los operadores WD de Yves Escoufier, transformados y normados (10).

Sexta etapa

La **sexta etapa** consitirá en cortar mediante una lineal horizontal o sinuosa, el **dendrograma** obtenido a partir de la **matriz de distancias,** construida en la **quinta etapa,** el **criterio de agregación de la varianza** y el **criterio concreto** que poseen los investigadores, conocedores de los **datos empíricos.** La finalidad de este proceso es la obtención de **clases homogéneas.**

Dado que en esta etapa concreta, para la construcción del **dendrograma,** no es necesario la utilización de un **algoritmo rápido de clasificación jerárquica ascendente,** puesto que la **matriz de distancias de partida** suele ser de dimensiones pequeñas en las encuestas de opinión (22), aconsejamos a los investigadores que apliquen el **algoritmo de clasificación jerárquica ascendente** basado en el **criterio de la varianza** y contemplado de forma didáctica (13) por Pierre Cazes (profeseur de la Université de Pierre-et-Marie-Curie, París, y miembro del comité científico de la Révue Les Cahiers de l'Analyse des Données).

Séptima etapa

La **séptima etapa** consistirá en la **diagonalización** de la **matriz T cuadrado de Tschuprow.** Esta fase nos permitirá no sólo la **representación gráfica** de las **variables cualitativas** en el **círculo de correlaciones,** sino también, cuando no todas las **variables cualitativas** presenten el mismo número de modalidades, nos apuntará que tendremos que hacer uso del **primer vector propio ortonormado,** extraído de la **matriz T cuadrado de Tschuprow,** para la aplicación de un **AFC,** tal como se contempla en (9).

Octava etapa

La **octava etapa** consistirá en la construcción de la **tabla de datos** para ser sometida a la aplicación de un **AFC.**

Se pueden presentar dos casos:

1.º Si el número de modalidades asociadas a todas las **variables cualitativas** es el mismo, la **tabla de datos** presentará la siguiente forma:

$$\left(\frac{B}{U}\right)$$

2.º Si el número de modalidades asociadas a todas las **variables cualitativas** no es el mismo, la **tabla de datos** presentará la siguiente forma:

$$\left(\frac{B^{o}}{U^{o}}\right)$$

Para no resultar reiterativos, el significado de las matrices contenidas en estas dos **tablas de datos** está contemplado en (9).

Novena etapa

La **novena etapa** consistirá en la aplicación de:

1.º Un **algoritmo de clasificación no-jerárquica** en el caso hipotético que la matriz de datos contenga, al menos, más de 200 filas (individuos).

2.º Un **algoritmo de clasificación jerárquica ascendente** en el caso hipotético de que la matriz de datos contenga, al menos, menos de 200 filas (individuos).

Observación de interés: Entre estos dos tipos de **algoritmos,** los que proporcionan **clases más estables** son los **algoritmos de clasificación jerárquica ascendente.**

1.º Aplicación de un algoritmo de clasificación no-jerárquica

Entre los mútiples **algoritmos de clasificación no-jerárquica** eligiremos el más adecuado, el cual especificaremos más adelante.

2.º Aplicación de un algoritmo de clasificación jerárquica ascendente

Una vez construidas las **clases estables,** obtenidas mediante la aplicación de un **algoritmo de clasificación no-jerárquica** a la **tabla de datos,** la **nueva tabla de datos** para la aplicación del **algoritmo de clasificación jerárquica ascendente** estará constituida —en esta ocasión— por los **baricentros** de las **clases estables.**

Observación de interés

En el caso hipotético de que tan sólo deseemos establecer una **tipología concreta** de los individuos y la **tabla de datos** no sea de grandes dimensiones, es decir, no exceda de 200 filas (individuos), aplicaremos un **algoritmo de clasificación jerárquica ascendente.**

En esta ocasión, la **tabla de datos** de partida contendrá las **coordenadas** de las filas (individuos) sobre los **ejes factoriales** más representativos, extraídos de la aplicación de un **AFC** (14, 15).

Décima etapa

La **décima etapa** consistirá en la utilización conjunta de los resultados obtenidos en la **octava** y **novena etapas.**

Recordemos que para tomar la decisión de cómo hay que cortar al **dendrograma** (línea horizontal o línea sinuosa), para la obtención de **clases homogéneas,** no sólo tenemos que hacer uso de estas dos etapas, sino también del **criterio concreto** que poseen los investigadores que conocen bien sus propios **datos empíricos** (16).

Así pues, esta forma de actuar no sólo está avalada por J. P. Benzécri —fundador y presidente de la revista **Les Cahiers de l'Analyse des Données** (publicada trimestralmente desde el año 1976) y profesor de la Université de Pierre-et-Marie-Curie, París—, sino también por el ilustre filósofo alemán Immanuel Kant, que dijo:

La teoría sin práctica es ciega.
La práctica sin teoría es absurda.

Undécima etapa

La **undécima etapa** consistirá en la aplicación del **análisis discriminante lineal** (AFDL) a una **tabla de datos.**

En última instancia hemos de indicar que cuando se obtengan las **clases homogéneas como resultado** de la **décima etapa,** es aconsejable la aplicación de un **análisis discriminante lineal.** Pues bien, siguiendo las recomendaciones —claramente explicadas— (17, 18), para que los resultados de dichos **análisis** sean fiables, hay que partir de una **tabla de datos** que contenga las **coordenadas de los individuos** sobre los **ejes factoriales** más significativos que provengan tanto de la aplicación de un **AFC** como la de un **ACM.** Recordemos que el número de **ejes factoriales** retenidos dependerá de las dimensiones de la **tabla de datos originales** y que como máximo serán 10.

Además, para que el **análisis discriminante lineal** tenga un cierto grado de fiabilidad, es recomendable trabajar con el 80% de los individuos de cada **clase** y dejar el resto con el fin de validar el **análisis.**

Observaciones referentes a las 11 etapas

De las **11 etapas** que constituye la **estrategia metodológica** basada en el **Análisis Estadístico Multidimensional Lineal, fuera de hipótesis distribucionales *a priori,*** tan sólo mostraremos las **etapas 3, 4, 5, 6, 7** y **8,** ya que las restantes son bastante conocidas por aquellas personas que manejen los principios del **Análisis Estadístico Multidimensional Lineal, fuera de hipótesis distribucionales *a priori.*** No obstante, para los investigadores que desconozcan la forma de proceder de las **etapas** que omitimos: **primera, segunda, décima y undécima,** les aconsejamos muy vivamente que consulten (19, 20, 21) y (14, 15), respectivamente.

Para llevar a cabo las **11 etapas** de la **estrategia metodológica** es necesario el establecimiento de unos **criterios de elección del programa o paquete de programas,** y después aconsejar cuáles son —a nuestro juicio— los más indicados en cada una de las **11 etapas** para llevar con **éxito** nuestros objetivos.

Criterios de elección del programa o paquetes de programas para la realización del estudio estadístico de una encuesta de opinión

En principio, los investigadores, entre la inmensidad de **métodos factoriales** y sobre todo de **algoritmos de clasificación** que están disponibles en el mercado, tal como se contemplan en (22), suelen estar perdidos, en numerosas ocasiones, y se les plantean numerosas dudas sobre cuál o cuáles deben **elegir,** para llevar a cabo con **éxito** sus propios análisis. Apuntamos, pues, que según sean sus objetivos, tendrán que hacer uso de unos u otros de sus objetivos.

Una observación en cuanto a los métodos factoriales

Siguiendo los consejos de J. Lemaire —Maître-Assistant à l'Université de Nice (1982)— en un Congreso impartido en Francia, recomendamos fuertemente que tanto

el programa como el paquete de programas que se **elija**, posea un **método factorial** que contenga los programas ya aludidos en (9) para el cálculo de los **valores propios** y de los **vectores propios** (3, 24, 25, 26, 27). Esto es debido a que el posicionamiento multidimensional —tanto de los **puntos-individuos activos** como de los **puntos-variables activos**— en los **planos factoriales,** va a depender de los programas que contenga el **método factorial,** para el cálculo de los **valores propios** y de los **vectores propios.**

Una observación en cuanto a los algoritmos de clasificación no-jerárquica

En cualquier programa o paquete de programas que contenga **algoritmos de clasificación no-jerárquica** debe estar implementado un **algoritmo** para la generación de números aleatorios.

J. P. Aselin de Beauville muestra en (28) un conjunto de programas para la generación de números pseudo-aleatorios. A pesar de esta indicación y de que existen numerosos programas, desde nuestro punto de vista, el que aconsejamos preferentemente es el **algoritmo Super-Duper** de Masaglia, basado en un generador multiplicativo y un generador de Taustworthe (29).

Al menos, que nosotros sepamos, el paquete de programas **S-PLUS** (versión 3.2, 1995) contiene este algoritmo.

A título informativo, hemos de indicar que el paquete de programas **S-PLUS** de origen americano tiene la misma estructura que el paquete de programas **STATlab** (versión 3.0, 1996) de origen francés. El responsable científico del **STATlab** es Michel Jambu —expertise et de savoir-faire dans le domaine des «Données client, du CRN et du Marketing—. Las tres modalidades de esta especialidad son:

1.ª Méthodes de Management de la Qualité de la Connaissance Client (CIO).

2.ª Customer Intelligence Learning (CIL).

3.ª Customer Intelligence Manager (CIM).

Hemos de decir que, aunque **S-PLUS** y **STATlab** son dos paquetes de programas excelentes para el **tratamiento estadístico de datos empíricos** provenientes de cierto tipo de investigaciones, no son los más indicados para el **tratamiento estadístico de las encuestas de opinión.**

En base a lo expuesto entre la inmensidad de paquetes de programas existentes en el mercado que contienen **métodos factoriales** y **algoritmos de clasificación** (22), tendríamos que aconsejar uno para el **tratamiento de las encuestas de opinión.** La pregunta que hay que hacerse es: ¿Qué programa o qué paquete de programas es el que nos proporciona mayor grado de credibilidad para que la aplicación a nuestros **datos empíricos** sea lo más fidedigna posible?

No está de más indicar que —a nuestro juicio— el paquete de programas más indicado para el **tratamiento estadístico de las encuestas de opinión** es la versión 7.0

de **SPAD.** A título informativo, esta versión se lanzó al mercado a finales del mes de noviembre de 2007. Así como el responsable científico del **STATlab** es Michel Jambu, el responsable científico del **SPAD** es Ludovic Lebart [École National Supérieur de Telécommunications (1995) y pertenece al comité científico de la revista «Les Cahiers de l'Analyse des Données»].

No obstante, aunque la **nueva estrategia metodológica** que proponemos no está contemplada en la nueva versión del paquete de programas **SPAD,** muchos de los programas que aconsejaremos en algunas de las **11 etapas,** sin duda, sí están integrados en **SPAD.**

Por otra parte, debe ser bien sabido por todos los autores que traten **datos empíricos,** ya sea provenientes —tanto de una **encuesta de opinión,** como de otro tipo de ensayos— que la **estrategia metodológica** que propongan debe estar constituida por dos **etapas** cuyo orden tienen que respetar:

1.º Etapa de estructuración de la información.

2.º Etapa de modelización.

En este artículo, para no extendernos demasiado, hemos abordado tan sólo la **etapa de estructuración de la información** adaptada al **tratamiento estadístico de las encuestas de opinión,** omitiendo la **etapa de modelización.**

Con el fin de **elegir** —con conocimiento de causa— cuál es el programa o paquete de programas que tenemos que utilizar —tanto en el **tratamiento estadístico de las encuestas de opinión** como en cualquier otro tipo de **datos empíricos—,** hemos elaborado un **decálogo.**

Decálogo para la elección del proprama o paquetes de programas

Acto seguido, en base a nuestra dilatada experiencia, entre la multitud de **métodos factoriales** y, sobre todo, de **algoritmos de clasificación,** nos tomaremos la libertad de aconsejar cuál o cuáles son los más adecuados, siendo fieles a un **decálogo** diseñado por nosotros:

Decálogo

1. **«CREDIBILIDAD». Completo en cuanto al número y calidad de prestaciones.**
2. **Conceptualmente profundo y riguroso.**
3. **Actualización:**
   - **3.1. Que esté actualizado.**
   - **3.2. Plasticidad (buena capacidad de modificación y posteriores actualizaciones).**

4. **Claridad expositiva (texto, representaciones gráficas).**
5. **Altura profesional.**
6. **Compatilibidad e interconexión fluida de intercambio con otros programas.**
7. **Rapidez de ejecución.**
8. **Distribución optimizada de sus contenidos.**
9. **Objetivos, presentes y futuros «claros».**
10. **Economía. A la vista de lo expuesto, estudiar —detenidamente— la relación «calidad/precio».**

En resumen: «Credibilidad —Facilidad de uso—, Economía».

Programas y paquetes de programas aconsejados en cada una de las 11 etapas de la estategia metodológica

Primera etapa

En la **primera etapa** de la **estrategia metodológica,** entre los **algoritmos de clasificación jerárquica descendente,** aconsejamos muy vivamente el propuesto por Williams y Lambert (19). Si bien, años más tarde, Michel Volle desarrolló este **algoritmo** de forma didáctica (20, 21).

No está de más recordar que este **algoritmo** nos permite reflejar en cada paso de la partición que realiza mediante su **criterio,** cuáles son las variables que aportan la **máxima información** con respecto a las restantes.

Segunda etapa

En la **segunda etapa** de la **estrategia metodológica** pensamos que no sólo es imprescindible la lectura del libro de J. P. Benzécri que —a nuestro juicio— es el pilar del **AFC** (30), sino que también merece la pena la de la monografía de Dervin (31), por su brillante exposición didáctica tanto del **AFC** como del **ACM.**

El programa del **AFC** no sólo está contenido en el paquete de programas **SPAD,** que cumple fielmente el apartado **6** del **decálogo** ya aludido para la **elección del programa o paquetes de programas,** sino también en otros paquetes de programas, tales como el **SICLA** y el **MODULAD.** Que nosotros sepamos, existe una interconexión entre el **SPAD, SICLA** y **MODULAD.** Por tanto, tanto el **AFC** como el **ACM** son los mismos.

Por otra parte, el paquete de programas **STATlab** contiene tanto el **AFC** como el **ACM.**

Otro paquete de programa de interés que no se contempla en (22) es el conocido con el nombre de **STATPC** (format WORD6 PC, 1997), cuyo responsable científico

es Thierry Foucart (Maître de conférences à l'Université de Poitiers y profesor de Estadística à l'IUT de Chatellerault). En dicho paquete de programas también se contempla el **AFC.**

Por último, existe otro paquete de programas conocido en el nombre de **STATIT-CF** (versión de 1990), desarrollado por un conjunto de investigadores pertenecientes a diversos centros de investigación. Aunque en dicho paquete de programas han intervenido muchos más investigarores de los que vamos a mencionar, citaremos alguno de estos, tales como a: C. Dervin (INRA-INA.-PG), J. P. Govet y G. Philippeau (SESI-ITCF), R. Tomassone (INA-PG) y R. Palm [Faculté des Sciences Agronomiques de Gembloux (Belgique)]. Ni que decir tiene que en este paquete de programas también se encuentra tanto el **AFC** como el **ACM.**

Observaciones de interés

Todos los paquetes de programas que hemos expuesto presentan una característica común: en todos ellos existen monografías, revistas o libros en los cuales se presentan —de forma didáctica—, el contenido estadístico de los programas.

En concreto, para el paquete de programas **SPAD,** aparte de una documentación detallada de su manejo, existen monografías que explican minuciosamente el contenido de los programas (32, 33).

Para el paquete de programas **SICLA** existe una documentación muy completa de sus contenidos (34, 35, 36, 37).

Para el paquete de programas **MODULAD,** además de una documentación detallada publicada en 1987 (38), desde el mes de junio de 1988 se ha estado editando una revista con periocidad semestral en la cual se explican los métodos estadísticos contenidos en **MODULAD** (38).

Para el paquete de programas **STATlab,** aparte de una documentación detallada de su manejo (39, 40, 41, 42), existe un libro escrito por Michel Jambu, que es muy didáctico sobre los **métodos factoriales** y los **algoritmos de clasificación** aplicando el **STATlab** (43).

Para el paquete de programas **STATPC** disponemos de un texto excelente —escrito por Thierry Foucart—, en el cual explica de forma didáctica no sólo los **métodos factoriales** sino también **los algoritmos de clasificación** aplicando el **STATPC.**

Con respecto al paquete de programas **STATIT-CF,** no sólo se ha desarrollado una documentación detallada de dicho paquete de programas, sino también muchas monografías de los programas contenidos en el mismo, entre ellas podemos destacar la dedicada al **AFC** y al **ACM** debida a Dervin (31).

Tercera etapa

En la **tercera etapa** de la **estrategia metodológica** desarrollaremos un programa que realice los cálculos de la forma más **rápida** posible. Dichos cálculos consistirán

en la construcción de las [p(p-1)/2] **tablas de contingencia** a partir de la **tabla disyuntiva completa,** que contendrá p **variables cualitativas.**

Cuarta etapa

En la **cuarta etapa** desarrollaremos un programa que realice los cálculos de la forma más **rápida** posible. Dichos cálculos consistirán en el cálculo de las [p(p-1)/2] **T cuadrado de Tschuprow,** a partir de las [p(p-1)/2] **tablas de contingencia.**

Quinta etapa

En este caso desarrollaremos un programa que realice los cálculos de la forma más **rápida** posible. Dichos cálculos consistirán en el cálculo de las [p(p-1)/2] distancias entre los **operadores WD** de Yves Escoufier, **transformados** (centrados) y **normados.**

Observación de interés

Las fórmulas para el cálculo de las distancias están contenidas en (10).

Sexta etapa

En esta etapa recomendamos muy especialmente el **algoritmo de clasificación jerárquica ascendente** expuesto —de forma didáctica— por Pierre Cazes —miembro del comité científico de la revista **Les Cahiers de l'Analyse des Données** y profesor de la Université de Pierre-et-Marie-Curie, París— (13), cuyo **criterio de agregación** es el de la **varianza.** Este **algoritmo** presenta propiedades muy interesantes respecto al resto y, además, dado que el orden de la **matriz de distancias** de partida es más bien bajo, consideramos que, entre el resto de los **algoritmos de clasificación jerárquica ascendente,** es el más adecuado para la construcción del **dendrograma,** uno de los elementos básicos para la contrucción de **clases homogéneas.**

Decimos que el **orden de la matriz de distancias** es más bien bajo, puesto que en un **cuestionario de opinión,** para que tenga cierto grado de credibilidad, el número de **variables cualitativas** no debe ser elevado (no más de 15 **variables cualitativas).**

Para aquellos autores que deseen tener más información a este respecto, les indicamos que lo consulten (22), pues se contempla en él un **cuestionario** sobre el **cierre de ventas.**

Dicho **cuestionario** diseñado —en inglés— por N. Rackham y adaptado —en castellano— por nosotros, tan sólo contiene 15 preguntas codificadas en una **escala de Likert.**

Séptima parte

Los programas de cálculo numérico, para la obtención de los **valores propios** y de los **vectores propios** de la **matriz T cuadrado de Tschuprow,** se encuentran —de forma aún más didáctica— en (27) en la bibliografía recomendada en (9).

En caso de no disponer de los programas para el cálculo de los **valores propios** y de los **vectores propios,** un programador con nociones de **cálculo numérico** podrá programarlos sin dificultad, haciendo uso (27) e implementarlos a nuestra **nueva estrategia metodológica** para el **tratamiento estadístico de las encuestas de opinión.**

Octava etapa

En la **octava etapa**, para la aplicación tanto del **AFC** como del **ACM,** utilizaremos los programas aconsejados en **segunda etapa.**

Novena etapa

En la **novena etapa** tendremos que hacer uso, tanto de los **algoritmos de clasificación jerárquica ascendente** como de los **algoritmos de clasificación no-jerárquica.**

Para la presentación de esta etapa adoptaremos dos **estrategias distintas,** dependiendo de las dimensiones de la **tabla de datos inicial.**

Primera estrategia [la tabla de datos contiene menos de 200 filas]

En este caso concreto, entre estos dos grandes tipos de **clasificaciones** retendremos los **algoritmos de clasificación jerárquica ascendentes.**

Observaciones de interés para la aplicación de un algoritmo de clasificación jerárquica ascendente

a) Nunca debe aplicarse un **algoritmo de clasificación jerárquica ascendente** a la **tabla de datos originales.**

Los motivos son los siguientes:

1. Si la **tabla de datos** contiene un número considerable de **variables,** aunque no sobrepase las 200 filas, la aplicación de dicho **algoritmo** conlleva un gasto elevado de memoria.

2. Además, si no reducimos el número de columnas de la **tabla de datos,** los resultados obtenidos de dicha aplicación no son fiables (14, 15).

b) No es aconsejable aplicar un **algoritmo de clasificación jerárquica ascendente** a una **tabla de datos** que contenga como columnas todas las **coorde-**

nadas de los **puntos-filas** (individuos) o de los **puntos-columnas** (variables) sobre los **ejes principales (**14, 15).

c) Es recomendable que la aplicación del **algoritmo de clasificación jerárquica ascendente** se realice sobre la **tabla de datos** de columnas que contengan las **coordenadas** de los **puntos-filas** (individuos) o **puntos-columnas** (variables), sobre los **ejes factoriales** que absorban un porcentaje de la **inercia aceptable** (14, 15).

d) El resultado final de un **algoritmo de clasificación jerárquica ascendente** conlleva a una **representación gráfica** conocida con el nombre de **dendrograma.**

Elección de un algoritmo de clasificación jerárquica ascendente

a) Cuando la **tabla disyuntiva completa** sea de unas dimensiones razonables y, entiéndase por razonables que las filas no superen a 200 y las columnas sean aún menos que las filas —situación que se puede solucionar llevando a cabo la estrategia contemplada en (7)—, aconsejamos la aplicación a los **datos empíricos** del **algoritmo de clasificación jerárquica rápida** propuesto por Maurice Roux —perteneciente al comité científico de la revista **Les Cahiers de l'Analyse des Données**— en (45).

Para la aplicación de este **algoritmo** es condición imprescindible que la **tabla de datos** sea **cuantitativa.** Así pues, la estrategia que proponemos es construir otra **tabla de datos** cuyas columnas estén construidas por las **coordenadas de los individuos** sobre no más de **10 ejes factoriales** extraídos de la aplicación de un **ACM.**

De esta manera, la **tabla de datos** ya es **cuantitativa** y, por tanto, se puede aplicar el **algoritmo de clasificación rápida** propuesto por Maurice Roux (45).

No está de más recordar las propiedades que presenta este **algoritmo** frente a los demás.

El **algoritmo CAHVAR** construye una jerarquía por agregaciones sucesivas, según el **criterio del momento de orden 2 de una partición.** La originalidad del **algoritmo** de Maurice Roux se basa en dos puntos fundamentales:

1.º Trabaja directamente con la **tabla de datos** y calcula las **distancias** a medida que las necesita.

2.º Utiliza el **algoritmo rápido** conocido con el nombre **«vecinos recíprocos».**

Que nosotros sepamos, **CAHVAR** se encuentra, al menos, en la versión 2.1 de **MODULAD.**

b) Cuando se trate de una **tabla de datos** que represente la **matriz de distancias** entre los **operadores WD,** de Yves Escoufier, **transformados** (centrados) y **normados,** cuyas dimensiones no es de gran tamaño [en el caso del **cierre de ventas** (22) es de (15, 15)], no es necesario hacer uso del **algoritmo rápido** de Maurice Roux.

En este caso, aconsejamos el **algoritmo** explicado —de forma didáctica— por el profesor Pierre Cazes en (13).

No está de más recordar que el **criterio de agregación** de dicho **algoritmo** es el de la **varianza.**

Segunda estrategia [la tabla de datos contiene más de 200 filas]

Para llevar a cabo esta **estrategia** haremos uso de la aplicación conjunta de un **algoritmo de clasificación no jerárquica** y de un **algoritmo de clasificación jerárquica ascendente.**

Cuando se trate de una **tabla de datos** que contenga sobre todo un gran número de **variables cualitativas,** el **proceso metodológico** aconsejado es el de reducir lo máximo posible las **variables cualitativas** aplicando la estrategia metodológica **contenida en** (7).

Aún así, puede ser que tengamos un número de **variables cualitativas** considerables, pero esta situación será mucho menor que la inicial. Esta situación tan sólo tendrá lugar cuando la **matriz de datos** de partida de la encuesta que se haga tenga grandes dimensiones. Ni que decir tiene que esta situación no debe contemplarse en las encuestas encaminadas para lanzar un producto al mercado. Que nosotros sepamos, aún no está comercializado el programa informático para resolver la metodología contemplada en (7), sino que además tan sólo existe una referencia sobre dicha metodología escrita en castellano (46). Recordamos a los lectores de este artículo que sepan programar y deseen llevar a cabo la ejecución de dicha metodología que consulten previamente (46).

Elección de un algoritmo de clasificación no-jerárquica

Para el **tratamiento estadístico** de una **tabla disyuntiva completa** que contenga, al menos, más de 200 filas y un número de **matrices de variables indicadoras,** cuyo número total de **variables indicadoras** no sea demasiado elevado, pensamos que lo mejor es que antes de aplicar el programa **CROMUL,** los investigadores consulten (22, 47). En (22), además de contener una traducción del francés al castellano de **CROMUL,** se explica —paso a paso—, los comandos que hay que pulsar para llegar al resultado final del mismo. Que nosotros sepamos, **CROMUL** no sólo está contenido en **SICLA,** sino también en la última versión 7.0 de **SPAD,** que salió a finales de noviembre de 2007.

El **algoritmo de clasificación cruzada CROMUL,** propuesto por Gérard Govaert (47) —professeur à l'IUT de Metz— presenta un mayor interés que el de Edwind Diday (48) —professeur à l'Université Paris IX Dauphine. Chef de projet à l'INRIA—, dado que obtiene simultáneamente, en base a un **criterio** y aplicando el **algoritmo de nubes dinámicas** de Edwind Diday (48), **clases simultáneas,** tanto para las filas como para las columnas.

Elección de un algoritmo de clasificación jerárquica ascendente

Con la **tabla de datos** construida por los **baricentros de las clases estables** obtenidas por la aplicación de **CROMUL,** aplicaremos el **algoritmo** que muestra Pierre Cazes en (13), dado que la **tabla de datos** será de grandes dimensiones.

Décima etapa

La **décima etapa** consistirá, no sólo en la aplicación conjunta del **método factorial** elegido en la **octava etapa,** y el **algoritmo de clasificación no-jerárquica y jerárquica** elegido en la **novena etapa,** sino también el **criterio** de los investigadores que conocen bien sus **datos empíricos.**

Mientras que para el estadístico, la **séptima etapa** le ayudará —parcialmente— a saber cómo tiene que realizar el corte al **dendrograma** (línea horizontal o línea sinuosa), la **octava etapa** le permitirá tomar una decisión más acertada. Así pues, desde un punto de vista orientativo, para tener una ligera idea de a qué nivel de agregación debe realizarse un corte al **dendrograma,** es necesario la aplicación conjunta de la **séptima** y **octava etapa.**

No obstante, en la decisión final de cómo debemos cortar el **dendrograma,** primará el **criterio** de los investigadores sobre el del estadístico. Esta prioridad la contempla J. P. Benzécri (16). De esta manera, en principio, se obtendrán **clases** lo más **homogéneas** posibles.

Undécima etapa

La **undécima etapa** consistirá en la aplicación de un programa de **análisis discriminante lineal** lo más actualizado posible, en cuanto a la **ayuda a la interpretación de los resultados.**

Tanto el programa **CAHMI** como el **DISC,** contenidos en la versión 2.1 de **MODULAD,** han sido reemplazados por el **CAHVAR** y el **DISC con dos modificaciones,** respectivamente. Las dos modificaciones del programa **DISC** han sido introducidas por Conchita Callant y Gilles Celeux —secrétaire de rédaction de la Revue Modulad— en la versión 2.2 de **MODULAD,** lanzada en 1989 y contenida en (49). Así como ya hemos mostrado las ventajas que presenta **CAHVAR** frente a **CAHMI,** no está de más mostrar las ventajas de **DISC con dos modificaciones,** frente a **DISC,** contenido en la versión 2.2 de **MODULAD.** Gilles Celeux y Jean-Christophe Turlot (50), basándose en el programa de Breiman y colaboradores contenido en (51), diseñan una regla de discriminación lineal para la **validación cruzada.**

Las dos modificaciones implementadas al programa **DISC** de la versión 2.1 funcionan ya en la versión 2.2 de **MODULAD:**

— Es posible validar la regla de decisión por la técnica de la **validación cruzada** contenida en (50). Esta posibilidad es particularmente útil cuando no disponemos de una muestra bastante grande para construir la **muestra test.**

— Es posible la introducción de probabilidades *a priori* de la aparición de clases a discriminar. En la versión 2.1, estas probabilidades *a priori,* eran implícitamente iguales.

Aplicación de la estrategia metodológica a un caso concreto

Con el objetivo de ilustrar —de forma didáctica— la **nueva estrategia metodológica,** basada en el **Análisis Estadístico Multidimensional, fuera de hipótesis distribucionales *a priori,*** hemos diseñado para su resolución un ejercicio de aplicación a una supuesta miniencuesta de opinión.

Ejercicio práctico

El Director General de una empresa de productos cosméticos ubicada en Francia, demanda al Departamento de Investigación de Mercados que le realice un estudio que contenga, no sólo un análisis de las variables más relevantes, sino también una tipología de sus empleados, en cuanto a su comportamiento humano. Para dar respuesta al Director General, el Director de Investigación de Mercados se pone en contacto con los sociólogos de su departamento para que diseñen un cuestionario asociado a este tema. Después de un estudio en profundidad, los sociólogos se deciden, con ayuda de los estadísticos, a diseñar un cuestionario —lo más adaptado posible— a los objetivos perseguidos.

Los estadísticos aconsejan a los sociólogos que para que el cuestionario tenga un grado de credibilidad aceptable, no debe contener más de quince preguntas y además, **como caso pitolo,** debe ser dirigido a un estrato de los empleados que tengan el mismo salario. Es decir, que se conserve el principio de la homogeneización del estrato y, por consiguiente, la muestra extraída será representativa.

Dado que se elige, como caso piloto, un estrato muy homogéneo, en un salario que contiene mucho personal, los estadísticos deciden realizar un muestreo aleatorio simple y llegan a la conclusión de que con 18 empleados es suficiente para obtener unos resultados fidedignos. Si el resultado de dicho análisis fuera satisfactorio, se extendería a todo el personal que se caracterizaría, en distintos estratos, según el nivel de salario; es decir, de categoría en la empresa. Esta situación llevaría consigo la realización de un muestreo estratificado y susequentemente, a la realización —en cada estrato— del estudio que se llevó a cabo en el primer caso.

Pues bien, después de un estudio en profundidad, los sociólogos decidieron retener cuatro preguntas de carácter dicotómico; es decir, si la respuesta es positiva, se le asigna un 1, y si la respuesta es negativa, se le asigna un 0.

Las preguntas fueron las siguientes:

V1. ¿Lee usted el periódico *Le Monde?*

V2. ¿Vive usted en el campo?

V3. ¿Ha votado usted en el último referéndum?

V4. ¿Utiliza usted medios anticonceptivos?

Los resultados de esta miniencuesta se reflejan en la tabla adjunta:

Tabla de datos lógica

	V1	V2	V3	V4
1	1	0	0	1
2	1	1	1	0
3	1	1	0	1
4	0	0	1	1
5	0	0	1	1
6	1	0	0	0
7	0	1	1	1
8	1	0	1	0
9	0	0	1	0
10	0	1	0	1
11	1	1	1	0
12	1	0	0	0
13	0	1	0	0
14	0	0	0	1
15	1	1	1	0
16	1	1	0	0
17	0	0	1	1
18	0	0	1	1

Observación: aunque estas mismas preguntas están contenidas en los apuntes y en el libro de Michel Volle (20, 21), el número de individuos que representan a los empleados de la empresa de un mismo nivel salarial es distinto.

A partir de esta **tabla de datos lógica,** el Director de Investigación de Mercados propone a sus estadísticos que apliquen el **Análisis Estadístico Multidimensional Lineal**, **fuera de hipótesis distribucionales *a priori*** a esta **tabla de datos lógica,** con el fin de responder a las preguntas concretas del Director Gerente de la empresa.

Para ello, los estadísticos, antes de aplicar la **nueva estrategia metodológica,** es necesario que realicen un minucioso **proceso de depuración** de la **tabla de datos lógica.**

1. Proceso de depuración de una tabla de datos lógica

1.1. *Eliminación de las filas y las columnas en las cuales todos los valores sean ceros*

Dado que en nuestro caso concreto todas las filas y las columnas, al menos, contienen un 1, la **tabla de datos lógica** permanece la misma. Por tanto, la **tabla de datos lógica** es:

Tabla de datos lógica

	V1	V2	V3	V4
1	1	0	0	1
2	1	1	1	0
3	1	1	0	1
4	0	0	1	1
5	0	0	1	1
6	1	0	0	0
7	0	1	1	1
8	1	0	1	0
9	0	0	1	0
10	0	1	0	1
11	1	1	1	0
12	1	0	0	0
13	0	1	0	0
14	0	0	0	1
15	1	1	1	0
16	1	1	0	0
17	0	0	1	1
18	0	0	1	1

1.2. *Si las marginales de las filas de la tabla de datos lógica no son iguales, es aconsejable desdoblar cada una de las columnas*

Dado que en nuestro caso concreto las marginales de las filas **no son iguales,** procederemos al desdoblamiento de las columnas obteniendo la siguiente **tabla de datos** (12).

Tabla de datos lógica desdoblada

	V1+	V2+	V3+	V4+	V1–	V2–	V3–	V4–
1	0	0	0	1	1	1	1	0
2	1	1	1	0	0	0	0	1
3	1	1	0	1	0	0	1	0
4	0	0	1	1	1	1	0	0
5	0	0	1	1	1	1	0	0
6	1	0	0	0	0	1	1	1
7	0	1	1	1	1	0	0	0
8	1	0	1	0	0	1	0	1
9	0	0	1	0	1	1	0	1
10	0	1	0	1	1	0	1	0
11	1	1	1	0	0	0	0	1
12	1	0	0	0	0	1	1	1
13	0	1	0	0	1	0	1	1
14	0	0	0	1	1	1	1	0
15	1	1	1	0	0	0	0	1
16	1	1	0	0	0	0	1	1
17	0	0	1	1	1	1	0	0
18	0	0	1	1	1	1	0	0

Observación: la **tabla de datos lógica desdoblada** presenta las mismas características de la **tabla de datos lógica** a nivel del proceso minucioso de **depuración.** Por consiguiente, retendremos dicha **tabla de datos** para ser sometida al **algoritmo de clasificación jerárquica descendente** de Willimas y Lambert (19, 20, 21).

2. Proceso de depuración de una tabla disyuntiva completa

Dado que la **tabla disyuntiva completa,** constituida por la **yuxtaposición vertical** de las cuatro **matrices de variables indicadoras,** asociadas a las cuatro **variables cualitativas** V1+, V2+, V3+ y V4+, es la misma que la **tabla disyuntiva completa** constituida por la **yuxtaposición vertical** de las cuatro **matrices de variables indicadoras,** asociadas a las **variables cualitativas** V1–, V2–, V3– y V4– a dos modalidades, para los **análisis,** retendremos tan sólo la **primera.** Por tanto, la **tabla disyuntiva completa,** que será sometida a los análisis que indicaremos más adelante, tendrá la siguiente forma:

Tabla disyuntiva completa

	V1+		V2+		V3+		V4+	
	V11+	V12+	V21+	V22+	V31+	V32+	V41+	V42+
	1	0	1	0	1	0	1	0
1	0	1	0	1	0	1	1	0
2	1	0	1	0	1	0	0	1
3	1	0	1	0	0	1	1	0
4	0	1	0	1	1	0	1	0
5	0	1	0	1	1	0	1	0
6	1	0	0	1	0	1	0	1
7	0	1	1	0	1	0	1	0
8	1	0	0	1	1	0	0	1
9	0	1	0	1	1	0	0	1
10	0	1	1	0	0	1	1	0
11	1	0	1	0	1	0	0	1
12	1	0	0	1	0	1	0	1
13	0	1	1	0	0	1	0	1
14	0	1	0	1	0	1	1	0
15	1	0	1	0	1	0	0	1
16	1	0	1	0	0	1	0	1
17	0	1	0	1	1	0	1	0
18	0	1	0	1	1	0	1	0

2.1. Eliminación de las columnas que no alcancen, al menos, un 2% de unos

Dado que en nuestro caso concreto, todas las columnas alcanzan cifras superiores al 2%, se conserva la misma **tabla de datos** (6).

2.2. Eliminación de alguna variable cualitativa

Dado que todas las variables cualitativas presentan, al menos, dos modalidades, la **tabla de datos** será la misma que en el apartado anterior.

Estrategia metodológica

La **estrategia metodológica** que vamos a mostrar está constituida por **11 etapas.**

Primera etapa

La **primera etapa** consistirá en la aplicación del **algoritmo de Williams** y **Lambert** (19), a una **tabla de datos lógica** y a la **tabla de datos lógica desdoblada,** asociada a las **variables indicadoras:** V1+, V2+, V3+, V4+, V1–, V2–, V3– y V4–, debidamente **depuradas.** Dado que este **algoritmo** se desarrolló en el año 1959, simplemente nos limitaremos a indicar —a los investigadores que lo desconozcan— cuál es la documentación más indicada en la que se encuentra una exposición didáctica de dicho **algoritmo** (19, 20, 21).

Segunda etapa

La **segunda etapa** consisitirá en la aplicación del **AFC** a la **tabla de datos lógica** y a la **tabla de datos lógica desdoblada** debidamente **depuradas.**

Así como la aplicación de un **AFC** a una **tabla de datos lógica** es un tema tan conocido como aplicado, la aplicación de un **AFC** a una **tabla de datos lógica desdoblada** es algo menos conocido, y por tal motivo se considera importante la consulta de la documentación (12).

De esta manera podrán realizar fácilmente, con un simple programa de **AFC,** estas dos aplicaciones.

Aunque en la actualidad es raro que en un paquete de programas no esté incluido el **AFC,** pensamos que podría —potencialmente— aportar aspectos muy interesantes y por lo tanto, aconsejamos la revisión de **SPAD.**

Tercera etapa

La **tercera etapa** consistirá en la construcción de **seis tablas de contingencia.**

Para la construción de las seis **tablas de contingencia** partiremos de la **tabla disyuntiva completa,** contenida en el **proceso de depuración de una tabla disyuntiva completa.**

Así pues, partiendo de la **tabla disyuntiva completa,** obtenemos sin dificultad las seis **tablas de contingencia.**

En nuestro caso concreto son las siguientes:

Tablas de contingencia

		V2+		
		1	0	
V1+	1	5	4	9
	0	3	6	9
		8	10	18

		V3+		
		1	0	
V1+	1	4	5	9
	0	6	3	9
		10	8	18

		V4+		
		1	0	
V1+	1	2	7	9
	0	7	2	9
		9	9	18

		V3+		
		1	0	
V2+	1	4	4	8
	0	6	4	10
		10	8	18

		V4+		
		1	0	
V2+	1	3	5	8
	0	6	4	10
		9	9	18

		V4+		
		1	0	
V3+	1	5	5	10
	0	4	4	8
		9	9	18

Cuarta etapa

La **cuarta etapa** consistirá en la construción de las seis **T cuadrado de Tschuprow,** obtenidas a partir de las **seis tablas de contingencia** obtenidas en la **tercera etapa.**

A partir de las **tablas de contingencia,** haciendo uso de la fórmula para el cálculo de la **T cuadrado de Tschuprow,** contenida en (9), obtenemos los siguientes resultados:

$$T^2 \text{ de Tschuprow}$$

$$T^2_{V1+,V2+} = 0,050000 \quad T^2_{V2+,V3+} = 0,010000$$
$$T^2_{V1+,V3+} = 0,050000 \quad T^2_{V2+,V4+} = 0,050000$$
$$T^2_{V1+,V4+} = 0,308642 \quad T^2_{V3+,V4+} = 0,000000$$

Por consiguiente, la matriz **T cuadrado de Tschuprow** adopta la siguiente forma:

$$\begin{pmatrix} 1,000000 & 0,050000 & 0,050000 & 0,308642 \\ 0,050000 & 1,000000 & 0,010000 & 0,050000 \\ 0,050000 & 0,010000 & 1,000000 & 0,000000 \\ 0,308642 & 0,050000 & 0,000000 & 1,000000 \end{pmatrix}$$

El cálculo de estos coeficientes son muy útiles porque nos permiten averiguar el grado de asociación existente entre las **variables cualitativas** V1+, V2+, V3+ y V4+ a dos modalidades.

Observación: por no resultar reiterativos cuando las **variables cualitativas** son: V1–, V2–, V3– y V4–, no hemos calculado la **matriz T cuadrada de Tschuprow,** ya que es la misma que la de las **variables cualitativas:** V1+, V2+, V3+ y V4+ resultados.

Quinta etapa

La **quinta etapa** consistirá en el cálculo de las distancias entre los **operadores WD** de Yves Escoufier, **transformados** (centrados) y **normados,** a partir de las seis **T cuadrado de Tschuprow,** calculadas en la **cuarta parte.**

Aplicando la fórmula para el cálculo de las **distancias entre los operadores WD** de Yves Escoufier, **transformados** (centrados) y **normados** contenida en (10), obtenemos —sin dificultad—, la **matriz de distancias.**

Dicha **matriz de distancias** es la que a continuación mostramos:

$$\begin{pmatrix} 0,000000 & 1,378405 & 1,378405 & 1,175889 \\ 1,378405 & 0,000000 & 1,407125 & 1,378405 \\ 1,378405 & 1,407125 & 0,000000 & 1,414214 \\ 1,175889 & 1,378405 & 1,414214 & 0,000000 \end{pmatrix}$$

Sexta etapa

La **sexta etapa** consistirá en la construcción de **clases homogéneas** mediante la aplicación del **criterio de la varianza** explicado por Pierre Cazes (13) a la **matriz de**

distancias entre los operadores de Yves Escoufier **transformados** (centrados) y **normados** y el propio **criterio** de los investigadores.

Invitamos a los investigadores que haciendo uso de (11, 13) y de su propio **criterio** construyan, en principio, **clases homogéneas.**

Séptima etapa

La **séptima etapa** la estructuraremos en dos partes.

1.ª Parte

La **primera parte** consistirá en la **diagonalización** de la **matriz T cuadrado de Tschuprow.**

2.ª Parte

La **segunda parte** consistirá en, a partir de los resultados de la **primera parte,** la **representación gráfica** de las seis **variables cualitativas** en los **círculos de correleciones** (1,2), (1,3), (1,4), (2,3), (2,4) y (3,4).

1.ª Parte

Partiendo de los programas para el cálculo de los **valores propios** y de los **vectores propios** ya mencionados en (9), procederemos a la aplicación de dichos programas a la **matriz T cuadrado de Tschuprow,** ya calculada en la **cuarta parte.**

En primer lugar calcularemos los **valores propios,** siendo éstos:

$$\lambda_1 \quad 1{,}3281938 \qquad \lambda_2 \quad 1{,}0002129$$
$$\lambda_3 \quad 0{,}9842606 \qquad \lambda_4 \quad 0{,}6873327$$

Acto seguido procederemos al cálculo de los **vectores propios ortonormados,** asociados a los **valores propios,** siendo éstos:

$$\vec{v}_1^{\,n} = \begin{pmatrix} 0{,}6907905 \\ 0{,}2125502 \\ 0{,}1117176 \\ 0{,}6820192 \end{pmatrix} \qquad \vec{v}_2^{\,n} = \begin{pmatrix} -0{,}0183890 \\ 0{,}1127618 \\ 0{,}9776195 \\ 0{,}1766540 \end{pmatrix}$$

$$\vec{v}_3^{\,n} = \begin{pmatrix} -0{,}1507520 \\ 0{,}9706195 \\ -0{,}1377810 \\ -0{,}1272320 \end{pmatrix} \qquad \vec{v}_4^{\,n} = \begin{pmatrix} -0{,}7069260 \\ -0{,}0022190 \\ 0{,}1131186 \\ 0{,}6981798 \end{pmatrix}$$

2.ª Parte

Ahora, a partir de las fórmulas contenidas en (9), estamos en condiciones de calcular las **coordenadas** de las **variables cualitativas:** V1+, V2+, V3+ y V4+, en los seis **círculos de correlaciones.**

Cálculo de las coordenadas de las variables V1+, V2+, V3+ y V4+ en los seis círculos de correlaciones

1-2	1	2	1-3	1	3
V1+	0,7961173	–0,0183910	V1+	0,7961173	–0,1495609
V2+	0,2449583	0,1127738	V2+	0,2449583	0,9629507
V3+	0,1287515	0,9777236	V3+	0,1287515	–0,1366924
V4+	0,7860086	–0,1766728	V4+	0,7860086	–0,1262268

1-4	1	4	2-3	2	3
V1+	0,7961173	–0,5860808	V1+	–0,0183910	–0,1495609
V2+	0,2449583	–0,0018397	V2+	0,1127738	0,9629507
V3+	0,1287515	0,0937816	V3+	0,9777236	–0,1366924
V4+	0,7860086	0,5788297	V4+	–0,1766728	–0,1262268

2-4	2	4	3-4	3	4
V1+	–0,0183910	–0,5860808	V1+	–0,1495609	–0,5860808
V2+	0,1127738	–0,0018397	V2+	0,9629507	–0,0018397
V3+	0,9777236	0,0937816	V3+	–0,1366924	0,0937816
V4+	–0,1766728	0,5788297	V4+	0,1262268	0,5788297

A partir de este momento, invitamos a los investigadores a que apliquen un programa de **representaciones gráficas** para que posicionen cada una de las cuatro **variables cualitativas:** V1+, V2+, V3+ y V4+ por un punto en el cual figure la **sigla** de cada **variable** en cada uno de los seis **círculos de correlaciones.**

No está de más recordar tres puntos de interés a la hora de **interpretar** las **variables cualitativas** proyectadas en los **círculos de corrrelaciones.**

1.º Todos los **puntos-variables** tienen que estar situados dentro de un **círculo** de radio la unidad.

2.º A medida que el extremo del **vector,** definido desde el origen de coordenadas del **círculo** —de radio la unidad— hasta el **punto-variable,** se aproxime al

círculo de correlaciones —de radio la unidad—, la **representatividad** de la **variable** irá aumentando. Se dice que dicha **variable** es totalmente **representativa** cuando se encuentra situada justo en la línea que delimita el **círculo de correlaciones.**

3.º La **correlación** entre las **variables** vendrá dada por el **coseno** del ángulo que forman los **vectores** que parten del origen de coordenadas hasta los **puntos-variables.**

4.º Los **puntos-variables** que se encuentren próximos del origen de coordenadas carecen de **interpretación.**

Observaciones de interés

Es totalmente lícito sumar **vectores** que partan desde el origen de coordenadas hasta el **punto-variable,** siempre y cuando éste no esté próximo al origen de coordenadas por lo dicho en 4.

Tal es así que es posible que dos **variables** sean poco **representativas** por separado, mientras que sumándolas sean mucho más **representativas.**

Octava etapa

La **octava etapa** consistirá en la construcción de la **tabla de datos** para ser sometida a la aplicación del **AFC.**

En nuestro caso concreto, la **tabla de datos** es la **yuxtaposición horizontal** de la **tabla de Burt** con la **tabla lógica,**

$$\left(\frac{B}{U}\right)$$

donde,

B es la tabla de Burt de dimensiones (8,8)

	V11+	V12+	V21+	V22+	V31+	V32+	V41+	V42+
V11+	9	0	5	4	4	5	2	7
V12+	0	9	3	6	6	3	7	2
V21+	5	3	8	0	4	4	3	5
V22+	4	6	0	10	6	4	6	4
V31+	4	6	4	6	10	0	5	5
V32+	5	3	4	4	0	8	4	4
V41+	2	7	3	6	5	4	9	0
V42+	7	2	5	4	5	4	0	9

U es la **tabla lógica** de dimensiones (18,8)

0001	0	1	0	1	0	1	1	0
0002	1	0	1	0	1	0	0	1
0003	1	0	1	0	0	1	1	0
0004	0	1	0	1	1	0	1	0
0005	0	1	0	1	1	0	1	0
0006	1	0	0	1	0	1	0	1
0007	0	1	1	0	1	0	1	0
0008	1	0	0	1	1	0	0	1
0009	0	1	0	1	1	0	0	1
0010	0	1	1	0	0	1	1	0
0011	1	0	1	0	1	0	0	1
0012	1	0	0	1	0	1	0	1
0013	0	1	1	0	0	1	0	1
0014	0	1	0	1	0	1	1	0
0015	1	0	1	0	1	0	0	1
0016	1	0	1	0	0	1	0	1
0017	0	1	0	1	1	0	1	0
0018	0	1	0	1	1	0	1	0

Así pues, la **tabla de datos** es:

$$\left(\frac{B}{U}\right)$$

Tal como hemos indicado en (9) para que el **AFC** de esta **tabla de datos** sea equivalente a la aplicación de un **AFCM,** tiene que verificarse según se indica en (52):

1.º Los **individuos** (filas) de la **tabla de Burt** se consideran como **individuos activos.**

2.º Los **individuos** (filas) de la **tabla lógica** se consideran como **individuos suplementarios.**

Invitamos —muy vivamente— a los investigadores a que apliquen no sólo la metodología adecuada a esta **tabla de datos,** sino también que, de entre los programas más aconsejados sobre el **AFC,** apliquen el que le proporcione mayor **ayuda a la interpretación de los resultados.**

Ni que decir tiene que este programa está contenido en la versión 7.0 de **SPAD.**

Como resultado de dicha aplicación, los investigadores dispondrán de una **representación simultánea** —tanto de los perfiles de frecuencias relativas de las filas como de las columnas—. Es decir, los **puntos** representados en los **planos factoriales** extraídos de un **AFC** representarán perfiles de frecuencias relativas.

Así como en un **ACP** no se pueden **interpretar** los **puntos-filas** y los **puntos-columnas** en un mismo **plano factorial,** en el **AFC** sí es posible porque, tanto las filas como las columnas **juegan un papel simétrico.**

Novena etapa

La **novena etapa** consistirá, en nuestro caso concreto, en la aplicación de un **algoritmo de clasificación jerárquica ascendente.** Tanto la **estrategia metodológica** como el programa aconsejado, ya lo hemos indicado con detalle con anterioridad. Así que recomendamos fuertemente a los investigadores que apliquen el programa **CAHVAR** a la **tabla de datos,** constituida por las **coordenadas de los puntos-individuos** sobre los **ejes factoriales** más representativos extraídos de un **ACM.**

Décima etapa

La **décima etapa** consistirá en la aplicación conjunta de la **octava** y **novena etapa,** más el **criterio** de los investigadores, que son los que conocen los **datos empíricos** con el fin de construir **clases** lo más **homogéneas** posibles.

Una vez que las **clases** estén perfectamente delimitadas, se procederá a pasar a la última etapa de la **estrategia metodológica.**

Undécima etapa

La **undécima etapa** consistirá en la aplicación de un **análisis discriminante lineal.** En nuestro caso concreto, ya que no partimos de un volumen de datos considerable, el programa **DISC** contenido en la versión 2.2 de **MODULAD** es el más indicado.

Observaciones de interés

— Cuando se aplica un **análisis discriminante lineal** a una **tabla de datos** que contenga las **coordenadas de los individuos** sobre los **ejes factoriales** más representativos (de 8 a 10 como máximo), la **distancia** de Mahalanobis se transforma en la **distancia** euclidiana clásica. De esta manera, la **interpretación** de los **puntos-individuos contenidos** en los planos factoriales provenientes de un **AFDL** se interpretan mejor.

— Cuando la **tabla de datos** contenga un gran número de observaciones (al menos, más de 200), se particionará en **clases** y de cada una de ellas se retendrá el 80% de las observaciones. De esta manera, podremos constatar si el 20% de las observaciones excluidas pertenecen o no a las **clases ya establecidas** en la **décima etapa.**

Un comentario sobre las dimensiones *a priori* de una tabla de datos para una encuesta de opinión

Según hemos experimentado, para la realización de un cuestionario de opinión, en principio, no debe contener más de 510 individuos y más de 15 variables cualitativas.

Aconsejamos que, si es posible, el número de modalidades asociadas a las variables cualitativas sea el mismo y no exceda de cinco.

Estas aseveracionas las hacemos en base a dos experiencias:

1.ª experiencia

Francisco Javier Díaz-Llanos, en el año 2002, planteaba un cuestionario sobre el **cierre de ventas** que contenía 15 preguntas retomadas del ingés y traducidas debidamente al castellano (22) por el libro escrito en 1987 por el profesor Neil Rackham titulado: «Making major sales».

Los dirijimos por carta a los jefes de ventas de 509 empresas y tan sólo nos respondieron 148.

2.ª experiencia

La Association National du Managemant des Achats confeccionó un cuestionario sobre las **estrategias de influencia en los centros de compra.** Este cuestionario contenía 500 empresas.

En esta ocasión, enviaron 500 formularios con una carta personalizada y 10 días después otra carta de recuerdo con otro cuestionario. En este caso concreto, la tasa de respuestas fue del 41%. De los 500 formularios tan sólo pudieron disponer de 187.

De lo que se desprende que la **tabla de datos** no va a contener más de 200 individuos y por tanto, es estos casos concretos, se procedería a la aplicación de un **algoritmo de clasificación jerárquica ascendente.**

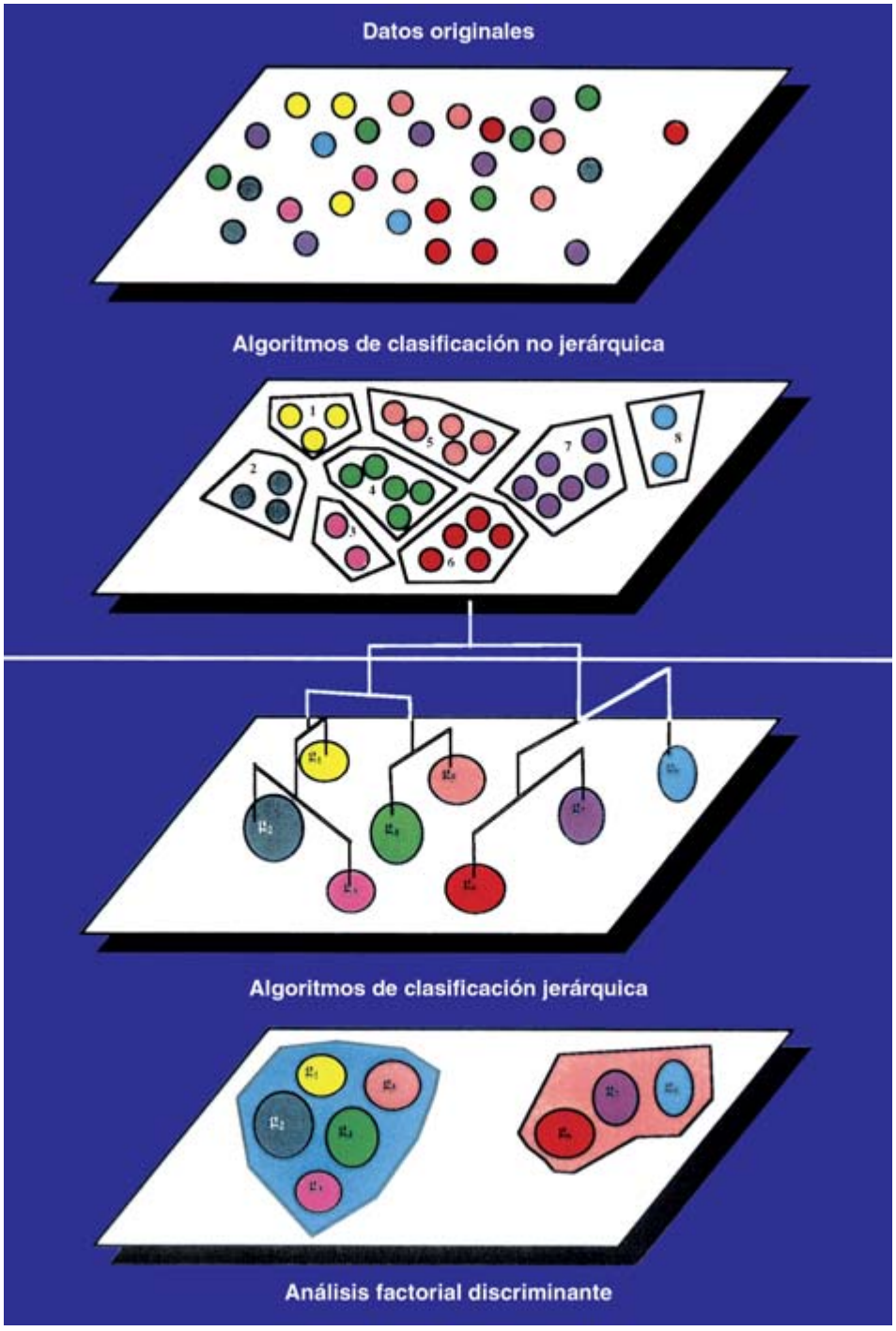
Conclusiones

Esperamos —muy vivamente— que esta **nueva estrategia metodológica,** que hemos propuesto, basada en el compromiso entre los **operadores WD** de Yves Escoufier y en técnicas actualizadas —tanto de **métodos factoriales** como de **algoritmos de clasificación**—, conlleve a una mejora, en la **ayuda a la interpretación de los resultados** en las encuestas de opinión.

Estudios posteriores de puesta a punto de este procedimiento desvelarán si dicho proceso conllevará a una **mejora en la ayuda a la interpretación de los resultados,** con respecto a los métodos tradicionales.

Agradecimientos

El grupo que ha realizado este estudio expresa su especial agradecimiento al profesor Yves Escoufier, sin cuya inestimable ayuda —técnica y moral—, no habría sido posible la realización de este trabajo. Asimismo, también agradecemos a todos sus colaboradores del Laboratorio de Biometría de Montpellier, la amistad y el asesoramiento prestado.


Datos originales
Algoritmos de clasificación no jerárquica
Algoritmos de clasificación jerárquica
Análisis factorial discriminante

BIBLIOGRAFÍA

(1) Dambroise, E. (1984). *Rapport de DEA. Etude d'opinion.* Université des Sciences et Techniques du Languedoc. Montpellier.

(2) Díaz-Llanos Sainz-Calleja, Fco. J. (1995). *El tratamiento estadístico de las encuestas de opinión, pieza clave en la Ingeniería de la demanda. Un enfoque didáctico y conceptual.* Ediciones CEES.

(3) Labart, L. (éditeur scientifique, editor) (1992). *La qualité de l'information dans les enquêtes*. Association pour la statistique et ses utilisations. Dunod.

(4) Brossier, G.; Dussaix, A.-M. (éditeurs scientifiques) (1999). *Enquêtes et sondages.* Dunod.

(5) Droesbeke, J.-J.; Lebart, L. (direction) (2001). *Enquêtes, modèles et applications.* Dunod.

(6) Benali, H. (1985). *Stabilité de l'analyse en composantes principales et de l'analyse de correspondances multiples en présence de certains types de perturbation. Méthodes de dépouillement d'enquêtes.* Thèse présenté devant l'Université de Rennes I. Spécialité: Traitement de l'Information. Option: Traitement Statistiques des Données.

(7) Dambroise, E.; Escoufier, Y.; Massotte, P. (1987). «Application de l'analyse des données à l'élaboration de minisondages d'opinion». *Révue de Statistique Appliquée,* le 11 mai, pp. 9-23.

(8) Díaz-Llanos Sainz-Calleja, Fco. J.; García Mouton, M. E. (1989). «Efecto y utilidad de la métrica en el análisis de encuestas». *Revista de Estadística Española.* Vol. 31, n.º 121, pp. 253-280.

(9) Díaz-Llanos Sainz-Calleja, Fco. J.; Tarazona Lafarga, J. V.; Valencia Delfa, J. L. (2007). «Efecto y utilidad del coeficiente RV de Yves Escoufier en el análisis de correspondencias múltiples». *Anales de la Real Academia de Doctores de España.* Vol. 11, n.º 1, pp. 9-44.

(10) Castilla Plaza, C.; Díaz-Llanos Sainz-Calleja, Fco. J.; Fernández Cancio, A. (2007). «Distancias entre los operadores WD de Yves Escoufier en el análisis de correspondencias múltiples». *Anales de la Real Academia de Doctores de España.* Vol. 11, n.º 1, pp. 45-64.

(11) Díaz-Llanos Sainz-Calleja, Fco. J.; Valencia Delfa, J. L. (2007). «Una CAH de los operadores WD de Yves Escoufier en el ACM». *Anales de la Real Academia de Doctores de España.* Vol. 12, n.º 2.

(12) Kobilinsky, A. (1975). *Presentation élémentaire de l'analyse factorielle des correspondances.* Département de Biométrie du CNRS. INRA (France).

(13) Maiti, D.; Thomas, Yves-F. (1975). *Interactions des plantes et du vent dans les dunes littorales.* Memoire du Laboratoire de Geomorphologie de l'École pratique des Hautes Études. F-35800 Dinard, n.º 28. En esta monografía ha colaborado el profesor Pierre Cazes.

(14) Roux, M. (1985). *Algorhitmes de classification.* Masson.

(15) Lebart, L.; Morineau, M.; Piron, N. (1995). *Statistique Exploratoire Multidimenssionnelle*. Dunod.

(16) Benzécri, F.; Benzécri, J. P., collaborateurs (1986). *Pratique de l'Analyse des Données en Économie*. Dunod.

(17) Wold, S. (1976). «Pattern recognitions by means of disjoint principal componet models». *Pattern Recognition,* 8, pp. 127-139.

(18) Benzécri, J. P. (1977a). «Analyse discriminante et analyse factorielle». *Les Cahiers de l'Analyse des Données,* 4, pp. 369-406.

(19) Lambert, J.; Williams, W. T. (1959). «Multivariate methods in plant ecology (1). Association analysis in plant communities». *J. Ecology,* 47, pp. 83-101.

(20) Volle, M. (1976). Institut National de la Statistique et des Etudes Economiques. Ecole National de la Statistique et de l'Administration Économique.

(21) Volle, M. (1980). *Analyse des données.* 2 ème édition. Economica.

(22) Díaz-Llanos Sainz-Calleja, Fco. J. (2002). *El análisis de datos en el cierre de ventas.* Editorial La Muralla, S. A. Hespérides, S. L.

(23) Householder, A. S. (1953). *Principles of Numerical Analysis.* McGraw-Hill, New York.

(24) Puy Huarte, J. (1983). *Cálculo numérico.* Apuntes de Cátedra de Matemática III de la ETSI de Caminos, Canales y Puertos, pp. 320-342, 399-402.

(25) Ciartet, P. G. (1985). *Introduction à l'analyse numerique matricielle et à l'optimisation*. Ed. Masson, pp. 90-94, 123-131.

(26) Ralston, A. (1978). *Introducción al análisis numérico.* Ed. Limusa. México.

(27) Burden, R. L.; Faires, J. D. (2004). *Métodos numéricos.* 3.ª edición. International Thomson Editores. Spain.

(28) Asselin de Beauville, J. P. (1974). «Les sous-programmes ussuels de simulation statistique». *Révue de Statistique Appliquée,* vol. XXII, n.º 4, pp. 57-87.

(29) Taustworthe, R. (1965). «Random number generated by linear recurrence modulo two». *Math Com,* pp. 19, 201-209.

(30) Benzécri, J. P & collaborateurs (1973). *L'Analyse des Données.* II. L'Analyse des Correspondances. Dunod.

(31) Dervin, C. (1990). *Comment interpréter les résultats d'une analyse factorielle des correspondances?* Institut Technique de Ceréales et de Fourrages.

(32) Aluja-Banet; Morineau, M. (2000). *Analyse en composantes principales (avec illustration SPAD).* CISIA. CERESTA Editeur.

(33) Morin, S.; Morineau, M. (2000). *Pratique du traitement des enquêtes. Exemples d'utilisation du Système SPAD.* CISIA. CERESTA Editeur.

(34) SICLA (1987). *Introduction à SICLA.* INRIA.

(35) Y. OK-Sakun (1987). *Le système SICLA. Manuel de l'utilisateur.* INRIA.

(36) SICLA (1989). *Les commamdes de SICLA*. INRIA.

(37) SICLA (1989). *Manuel de mise en oeuvre et exemples.* INRIA.

(38) MODULAD (1987). *Biblioteque Fortran 77 pour l'Analyse des données.* Version 2.1. Edité par l'Institut National de Recherche en Informatique et en Automatique. Domaine de Voluceau-Rocquencourt. BP 105.78153 Le Chesnay Cedex France.

(39) SLP Infoware (1996). *STATlab by SLP. The software for exploratory Data Analysis.* Users'guide.

(40) SLP Infoware (1996). *STATlab by SLP. Getting Started.* Traducido al castellano por Fco. J. Díaz-Llanos Sainz-Calleja.

(41) STATlab. Versión 3.0 (1996). CNET-France Telecom-SLP.

(42) Campos Sánchez, L.; Díaz-Llanos Sainz-Calleja, Fco. J. (1997). *Procedimientos de gestión informática utilizando el STATlab y sus aplicaciones en la Estadística Exploratoria Multidimensional.* Oficina Provincial del Registro de la Propiedad Intelectual, solicitud n.º 63787, Madrid.

(43) Jambu, M. (1999). *Méthodes de base de l'analyse des données.* Éditions Eyrolles et Frace Telecom-CNET.

(44) Foucart, Th. (1997). *L'analyse des données. Mode d'amploi.* Presses Universitaires de Rennes.

(45) Rox, M. (1989). «Construction ascendente hiérarchique rapide. Le programme CAHVAR». *Révue Modulad,* 3, 1-6.

(46) García Mouton, M. E. (1989). *Concurso de acceso al Cuerpo de Profesores Titulares de Universidad.* Proyecto docente y de investigación. Área de conocimientos: Matemática Aplicada. Departamento: Matemática Aplicada a la Ingeniería Agronómica. Universidad Politécnica de Madrid. ETSIA.

(47) Govaert, G. (1983). *Classification croisée*. Thèse de Doctoral d'État ès Sciences Mathématiques presenté à l'Université Pierre et Marie Curie. Paris VI.

(48) Diday, E. (1970). «La méthode des nuées dynamiques et la reconnaissance de formes». *Révue de Statistique Appliquée,* 19, 2.

(49) Callant, C.; Celeux, G. (1990). «Nouvelle version de DISC». *La Révue de Modulad. Périodique semestral du Club Modulad,* n.º 5, pp. 49-51. Édité par l'INRIA.

(50) Celeux, G.; Turlot, J.-Chr. (1989). «Estimation de la qualité d'une règle discriminante». *La Révue de Modulad. Périodique semestral du Club Modulad,* n.º 4, pp. 37-46. Édité par l'INRIA.

(51) Breiman, L.; Friedman, J. H.; Ohlson, R. A.; Stone, C. J. (1984). *Classification regressión trees.* Wadsworh.

(52) Diday, E.; Lemaire, J.; Pouget, J.; Testu, J. (1982). *Éléments d'analyse des données*. Dunod.