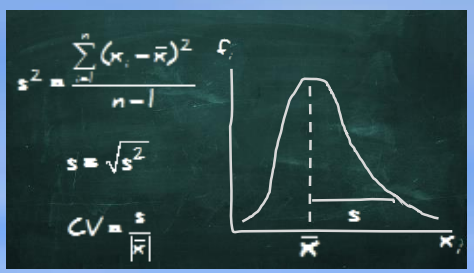
METODOLOGÍA DE LA INVESTIGACIÓN SOCIAL CUANTITATIVA

Pedro López-Roldán
Sandra Fachelli



UAB
Universitat Autònoma de Barcelona

METODOLOGÍA DE LA INVESTIGACIÓN SOCIAL CUANTITATIVA

Pedro López-Roldán
Sandra Fachelli

Bellaterra (Cerdanyola del Vallès) | Barcelona
Dipòsit Digital de Documents
Universitat Autònoma de Barcelona

Pedro López-Roldán
Centre d'Estudis Sociològics sobre la Vida Quotidiana i el Treball (http://quit.uab.cat)
Institut d'Estudis del Treball (http://iet.uab.cat/)
Departament de Sociologia. Universitat Autònoma de Barcelona
pedro.lopez.roldan@uab.cat

Sandra Fachelli
Departament de Sociologia i Anàlisi de les Organitzacions
Universitat de Barcelona
Grup de Recerca en Eduació i Treball (http://grupsderecerca.uab.cat/gret)
Departament de Sociologia. Universitat Autònoma de Barcelona
sandra.fachelli@ub.edu

Edición digital: http://ddd.uab.cat/record/129382

1ª edición, febrero de 2015

Edifici B · Campus de la UAB · 08193 Bellaterra
(Cerdanyola del Vallés) · Barcelona · España
Tel. +34 93 581 1676

Índice general

PRESENTACIÓN

PARTE I. METODOLOGÍA

I.1. FUNDAMENTOS METODOLÓGICOS
I.2. EL PROCESO DE INVESTIGACIÓN
I.3. PERSPECTIVAS METODOLÓGICAS Y DISEÑOS MIXTOS
I.4. CLASIFICACIÓN DE LAS TÉCNICAS DE INVESTIGACIÓN

PARTE II. PRODUCCIÓN

II.1. LA MEDICIÓN DE LOS FENÓMENOS SOCIALES
II.2. FUENTES DE DATOS
II.3. EL MÉTODO DE LA ENCUESTA SOCIAL
II.4. EL DISEÑO DE LA MUESTRA
II.5. LA INVESTIGACIÓN EXPERIMENTAL

PARTE III. ANÁLISIS

III.1. SOFTWARE PARA EL ANÁLISIS DE DATOS: SPSS, R Y SPAD
III.2. PREPARACIÓN DE LOS DATOS PARA EL ANÁLISIS
III.3. ANÁLISIS DESCRIPTIVO DE DATOS CON UNA VARIABLE
III.4. FUNDAMENTOS DE ESTADÍSTICA INFERENCIAL
III.5. CLASIFICACIÓN DE LAS TÉCNICAS DE ANÁLISIS DE DATOS
III.6. ANÁLISIS DE TABLAS DE CONTINGENCIA
III.7. ANÁLISIS LOG-LINEAL
III.8. ANÁLISIS DE VARIANZA
III.9. ANÁLISIS DE REGRESIÓN
III.10. ANÁLISIS DE REGRESIÓN LOGÍSTICA
III.11. ANÁLISIS FACTORIAL
III.12. ANÁLISIS DE CLASIFICACIÓN

Metodología de la Investigación Social Cuantitativa

Pedro López-Roldán
Sandra Fachelli

PARTE III. ANÁLISIS

Capítulo III.12
Análisis de clasificación

Bellaterra (Cerdanyola del Vallès) | Barcelona
Dipòsit Digital de Documents
Universitat Autònoma de Barcelona

Cómo citar este capítulo:

López-Roldán, P.; Fachelli, S. (2015). Análisis de clasificación. En P. López-Roldán y S. Fachelli, *Metodología de la Investigación Social Cuantitativa*. Bellaterra (Cerdanyola del Vallès): Dipòsit Digital de Documents, Universitat Autònoma de Barcelona. 1ª edición. Edición digital: http://ddd.uab.cat/record/142929

Capítulo acabado de redactar en agosto de 2015

Índice de contenidos

1. CLASIFICACIÓN, TIPOLOGÍA Y TAXONOMÍA. 6
2. CARACTERÍSTICAS, OBJETIVOS Y MODELO DE ANÁLISIS 10
   2.1. Etapa 1: Selección de las variables 13
   2.2. Etapa 2: Elección de la medida de proximidad 14
      2.2.1. Medidas de proximidad para variables continuas 14
      2.2.2. Medidas de proximidad para variables binarias 17
      2.2.3. Medidas de proximidad para datos de frecuencias 19
      2.2.4. La matriz de distancias 19
   2.3. Etapa 3: Elección del método de clasificación 20
      2.3.1. Métodos jerárquicos 22
         2.3.1.1. Método de distancias mínimas 25
         2.3.1.2. Método de distancias máximas 29
         2.3.1.3. Método de la distancia media entre grupos 30
         2.3.1.4. Método de la distancia media intra grupos 32
         2.3.1.5. Método de las distancias medianas 33
         2.3.1.6. Método de las distancias entre centroides 35
         2.3.1.7. El método Ward 35
      2.3.2. Métodos no jerárquicos 40
         2.3.2.1. Centros móviles 40
         2.3.2.2. Método de grupos estables 41
      2.3.3. Métodos mixtos 42
         2.3.3.1. Clasificación híbrida 42
         2.3.3.2. Clasificación en dos fases 43
   2.4. Etapa 4: Clasificación y número de grupos 43
   2.5. Etapa 5: Validación e interpretación de los resultados 45
3. EJEMPLOS DE APLICACIÓN 46
   3.1. Estratificación censal de la Región Metropolitana de Barcelona 46
   3.2. La conformación del perfil socio-espacial de la Ciudad de Buenos Aires 57
   3.3. Un modelo de estratificación para Argentina: los estratos sociales 60
4. ANÁLISIS DE CLASIFICACIÓN CON SPSS 64
5. ANÁLISIS DE CLASIFICACIÓN CON SPAD 74
6. ANÁLISIS DE CLASIFICACIÓN CON R 94
7. BIBLIOGRAFÍA 95

PARTE III
Capítulo 12

Análisis de clasificación

Una de las aproximaciones metodológicas, con lo que implica en métodos y técnicas, más frecuentes y características de la investigación científica en general, y de las ciencias sociales en particular, es el recurso a la construcción de tipologías, de clasificaciones, como ordenadoras de las distintas conceptualizaciones de los fenómenos sociales complejos estudiados. La construcción de tipologías satisface la necesidad de clasificar o de estructurar y, en general, de resumir en un conjunto reducido y significativo de categorías o tipos a los individuos, grupos, instituciones, sociedades o cualquier otra unidad de análisis que es objeto de estudio. Constituye uno de los procedimientos y uno de los objetivos más habituales de la investigación empírica sociológica desde los orígenes mismos de la disciplina.

La tipología ha sido objeto de atención habitual en diversas disciplinas. En la tradición sociológica en particular ha sido recurrente su presencia en los autores clásicos y en diversas propuestas metodológicas, con diversidad de perspectivas. Buena parte de las teorías sociológicas han encontrado a la tipología la forma de ordenar conceptualmente los fenómenos más diversos de la realidad social. Menos frecuentes han sido las contribuciones metodológicas de reflexión sobre su naturaleza y sobre los procedimientos de construcción. Esta constatación es especialmente significativa en los últimos tiempos cuando se recurre a las técnicas estadísticas la extensión de las que se ha potenciado con la difusión de la informática aplicada. Un proceso acelerado de posibilidades tecnológicas no siempre se ve acompañado de una preocupación por las necesidades metodológicas que requiere la investigación y poder llenar así su función de vincular la teorización de un problema sociológico con los instrumentos que permitan la contrastación de modelos teóricos y de hipótesis. Las técnicas de clasificación automática destinadas a la construcción de tipologías pueden caer en un exceso de empirismo que hay que anunciar desde un principio.

En este capítulo introduciremos una visión panorámica del conjunto de procedimientos de análisis multivariable que de manera genérica se denominan como técnicas de análisis de clasificación (automática) que se suele identificar en inglés como *cluster analysis* y cuya traducción se correspondería literalmente con la expresión análisis de conglomerados. El análisis de clasificación, como técnica estadística, está destinada al análisis del universo de los individuos, de las relaciones entre los casos de la matriz. En general, el objetivo consiste en la formación de grupos o clases de individuos homogéneos (similares o próximos) según las características (sociales) que definen un

conjunto variables que actúan de criterios de clasificación. Además, como sugerimos al tema anterior, estos procedimientos clasificatorios se pueden utilizar de forma complementaria con un análisis factorial.

Antes de presentar estas técnicas de clasificación, no obstante, precisaremos al inicio del capítulo tres conceptos similares que se utilizan habitualmente en el lenguaje cotidiano y científico y que requieren una clarificación; se trata de diferenciar el concepto de clasificación del de tipología y del de taxonomía. El capítulo continúa con la exposición de las características básicas, los objetivos y del modelo de análisis que identifican el análisis de clasificación, para terminar con la presentación de las diferentes etapas del procedimiento general de análisis destinado a la elaboración de una clasificación.

En el desarrollo de estos contenidos retomaremos algunos ejemplos utilizados en el capítulo anterior pues en ellos se ha planteado una metodología general destinada a la construcción tipológica donde se concibe un diseño de análisis que combina el análisis factorial con el análisis de clasificación. Este diseño es adecuado para desarrollar el modelo metodológico que denominamos de construcción de una tipología estructural y articulada (López-Roldán, 1994, 1996a; López-Roldán y Fachelli, 2015).

1. Clasificación, tipología y taxonomía.

Una de las cuestiones que surge cuando se utilizan los procedimientos utilizados y las reflexiones realizadas sobre el uso de clasificaciones y de las tipologías es la distinción entre contenido y forma. Una clasificación, una tipología, como contenido refleja la naturaleza sustantiva y diversa de los fenómenos sociales: según determinados criterios sustantivos se pueden encontrar diferentes tipos de sociedades, tipos de acción social, tipo de capital, tipo de empleo, tipos de estados del bienestar, de clases sociales, etc. La tipología como forma es una abstracción expresada en términos clasificatorios que nos permite afirmar que las sociedades, la acción social, el capital, el empleo, los estados del bienestar o las clases sociales, se pueden reconocer y clasificar a través de una diversidad de tipos.

Desde un punto de vista formal además puede introducirse una mayor precisión al distinguir clasificación de tipología, y éstas de otro concepto cercano, el de taxonomía, que en el lenguaje científico a veces se emplean indistintamente. Utilizadas de manera genérica, clasificación o tipología aluden a la habitual tarea de ordenación y reducción de los fenómenos o unidades que son estudiadas en un número limitado de categorías. El concepto de tipología y, en particular, el de tipo tiene una acepción con la que se alude a alguna noción que resume una diversidad de características, situaciones, fenómenos o individuos que comparten algún carácter más evidente o notorio y que pueden identificarse como modelo o prototipo diferenciado. Así se suele utilizar de forma habitual en el lenguaje cotidiano y también en el lenguaje científico. En muchos sentidos coincide con el concepto de clase y de clasificación ya que tienen un significado más global y genérico, pero que es preciso distinguir, y también del de taxonomía.

A esta tarea de precisión conceptual dedica un artículo Marradi (1990), donde se analizan los conceptos de clasificación, de tipología y de taxonomía para mostrar el

papel de las actividades clasificatorias en el trabajo científico. El análisis del autor parte de una primera distinción fundamental entre lo que son las operaciones de clasificación y los productos derivados de la actividad clasificatoria. Según Marradi, el término clasificación es habitualmente empleado para identificar tres tipos de operaciones distintas.

En primer lugar, se encuentran las llamadas clasificaciones intensivas que consisten en una operación de carácter intelectual donde la extensión de un concepto (genus) se subdivide en dos o más extensiones (species) a un nivel menor de generalidad de acuerdo con uno o varios criterios de división (fundamentum o fundamenta divisionis). Esta operación implica básicamente un proceso de elaboración conceptual donde la intensión o intensidad del concepto articula y clarifica en sus extensiones antes de ser reconocidas en la realidad empírica.

En segundo lugar, se distinguen las operaciones denominadas clasificaciones extensivas, donde los objetos o fenómenos de un conjunto dado se agrupan en dos o más subconjuntos según las similitudes derivadas de una o varias propiedades. Este tipo de operaciones son características de los análisis que parten de una matriz de datos, vectores de objetos/fenómenos cuyas componentes son las variables o propiedades definidas operacionalmente. En este caso, hasta que los grupos no son formados por algún procedimiento, no se establece el concepto que unifica cada combinación particular de elementos en el grupo constituido. En este sentido se expresan las técnicas de clasificación automática.

Por último, el tercer tipo de operación clasificatoria es identificada por el acto de asignar los objetos/fenómenos a las distintas categorías que previamente se han establecido[1]. Por tanto, esta operación es posterior a la definición de las categorías que se establecen a través tanto de la primera como de la segunda operación clasificatoria. Esta operación, además, puede aplicarse a objetos/fenómenos adicionales que no pertenecen al conjunto original que ha servido de base a la operación clasificatoria.

Esta distinción en tres categorías de operaciones sintetiza las principales familias de sentidos que tiene la operación clasificatoria y pueden dar lugar a diferentes resultados o productos: cuando se considera una subdivisión o una extensión de un solo concepto, el producto obtenido es una lista de clases, un esquema clasificatorio o simplemente una clasificación, compuesta de varias categorías denominadas clases; cuando se opera a la vez con varios principios clasificatorios tratados simultáneamente obtenemos una tipología, con varios tipos, y cuando estos principios clasificatorios se consideran como criterios sucesivos y jerárquicos, el orden clasificatorio resultante de la operación es una taxonomía que incluye varios taxones.

Disponemos pues de conceptos diferentes que las técnicas de análisis multivariable de clasificación no incorporan de forma clara ni unívoca. La literatura que trata de estos

[1] Es de interés destacar que en castellano o en catalán, por ejemplo, se emplea el término «clasificar» o «classificar» para hacer referencia a dos tipos de conceptos distintos que a la lengua inglesa o francesa identifican con palabras diferentes. Por un lado, clasificar es entendido como el acto de concebir, de formar, de ordenar o de dividir en clases. Esta idea se llama por el verbo inglés *to classsify* o el francés *classifier*. Por el otro, clasificar se corresponde con la acción de repartir o de asignar según una clasificación, para lo cual se emplean los términos *to class* en inglés y *classer* en francés. Esta distinción es la base que diferencia la primera y la tercera operación clasificatoria que comentamos de Marradi.

procedimientos ha generado una terminología rica y diversa que en ocasiones provoca la utilización contradictoria de algunas expresiones. En todo caso, en relación a la distinción que acabamos de establecer, las técnicas de análisis multivariable sólo podrán proporcionarnos información multidimensional, por tanto, generarán tipologías o taxonomías, no clasificaciones pues tienen un carácter de unidimensionalidad. Contradictoriamente, estas técnicas se denominan de clasificación. En este sentido la expresión clasificación la identificaremos de forma genérica con la operación de formación de clases y la asignación de las unidades en cada una de ellas. En un sentido más estricto, o en el sentido que hemos comentado a partir del análisis de Marradi, una clasificación es la expresión desglosada de un concepto único, mientras que la tipología lo es de un conjunto de conceptos que se combinan.

Las técnicas multivariables de clasificación deben entenderse como la operación general de constitución de grupos o clases. En la tradición francesa se emplean a menudo de manera indistinta las expresiones analyse de clasificación automatique o analyse tipologique para identificar esta técnica de análisis multivariable. En la tradición anglosajona la expresión es de cluster analysis. Aquí utilizaremos la expresión técnicas de clasificación (automática) para designar a la técnica de análisis multivariable, y preferimos utilizar un término como análisis tipológico para aludir a un proceso metodológico más general destinado a la construcción de tipologías, profuso en la tradición sociológica, que eventualmente puede usar técnicas de tipo multivariable.

Gráfico III.12.1. Ejemplos de clasificaciones, taxonomías y tipologías

1) Clasificaciones

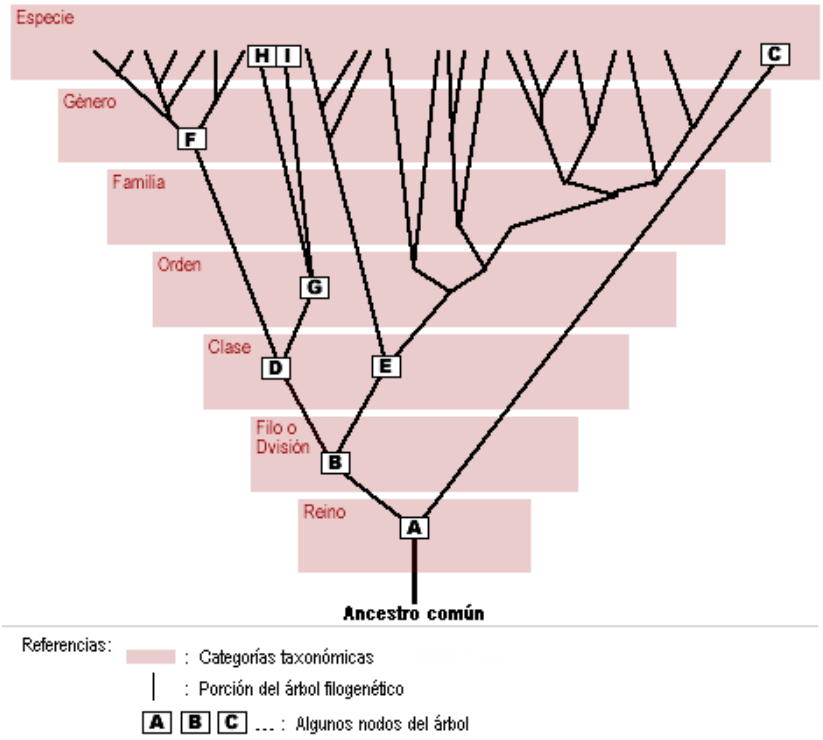
- CNAE, Clasificación Nacional de Actividades Económicas
  http://www.ine.es/jaxi/menu.do?type=pcaxis&path=/t40/clasrev&file=inebase
- CNO, Clasificación Nacional de Ocupaciones
  http://www.ine.es/jaxi/menu.do?type=pcaxis&path=/t40/cno11&file=inebase

1	**Directores y gerentes**
A	***Directores y gerentes***
11	Miembros del poder ejecutivo y de los cuerpos legislativos; directivos de la Administración Pública y organizaciones de interés social; directores ejecutivos
111	**Miembros del poder ejecutivo y de los cuerpos legislativos; directivos de la Administración Pública y organizaciones de interés social**
1111	Miembros del poder ejecutivo (nacional, autonómico y local) y del poder legislativo
1112	Personal directivo de la Administración Pública
1113	Directores de organizaciones de interés social
112	**Directores generales y presidentes ejecutivos**
1120	Directores generales y presidentes ejecutivos
12	Directores de departamentos administrativos y comerciales
121	**Directores de departamentos administrativos**
1211	Directores financieros
1212	Directores de recursos humanos
1219	Directores de políticas y planificación y de otros departamentos administrativos no clasificados bajo otros epígrafes
122	**Directores comerciales, de publicidad, relaciones públicas y de investigación y desarrollo**
1221	Directores comerciales y de ventas
1222	Directores de publicidad y relaciones públicas
1223	Directores de investigación y desarrollo
13	Directores de producción y operaciones
131	**Directores de producción de explotaciones agropecuarias, forestales y pesqueras, y de industrias manufactureras, de minería, construcción y distribución**
1311	Directores de producción de explotaciones agropecuarias y forestales
1312	Directores de producción de explotaciones pesqueras y acuícolas
1313	Directores de industrias manufactureras
1314	Directores de explotaciones mineras
1315	Directores de empresas de abastecimiento, transporte, distribución y afines
1316	Directores de empresas de construcción
132	**Directores de servicios de tecnologías de la información y las comunicaciones (TIC) y de empresas de servicios profesionales**
1321	Directores de servicios de tecnologías de la información y las comunicaciones (TIC)
1322	Directores de servicios sociales para niños
1323	Directores-gerentes de centros sanitarios
1324	Directores de servicios sociales para personas mayores

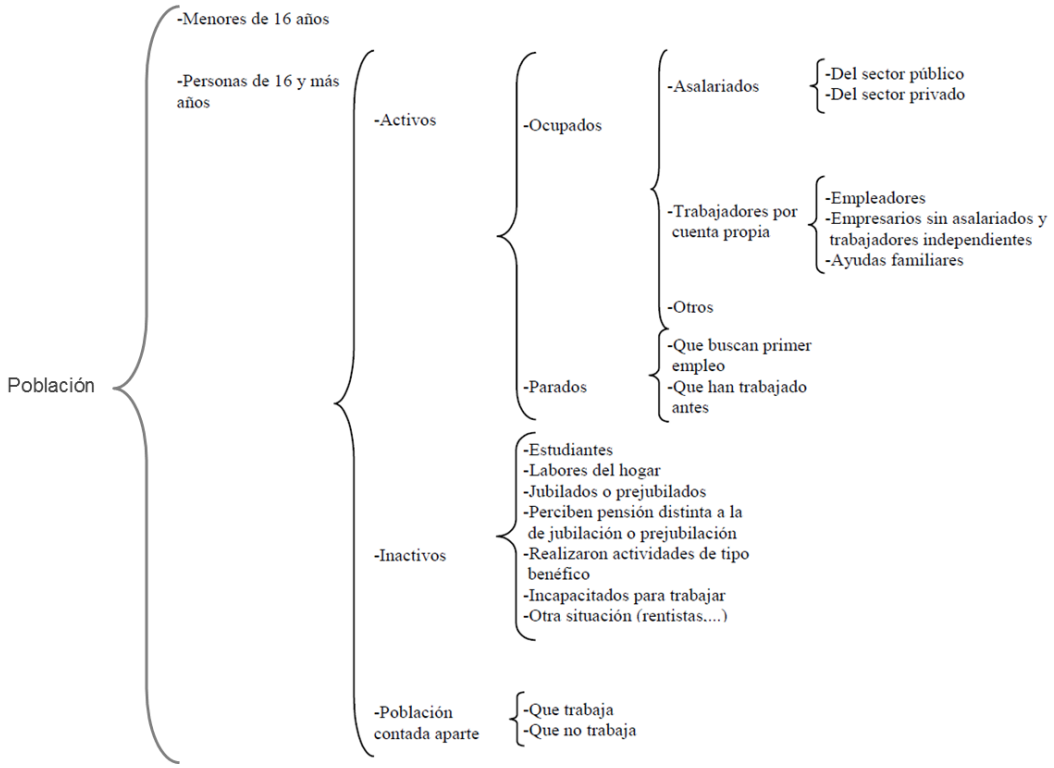
1	Directores y gerentes
2	Técnicos y profesionales científicos e intelectuales
3	Técnicos; profesionales de apoyo
4	Empleados contables, administrativos y otros empleados de oficina
5	Trabajadores de los servicios de restauración, personales, protección y vendedores
6	Trabajadores cualificados en el sector agrícola, ganadero, forestal y pesquero
7	Artesanos y trabajadores cualificados de las industrias manufactureras y la construcción (excepto operadores de instalaciones y maquinaria)
8	Operadores de instalaciones y maquinaria, y montadores
9	Ocupaciones elementales
0	Ocupaciones militares

2) Taxonomías

- Clasificaciones taxonómicas de minerales, vegetales y animales:

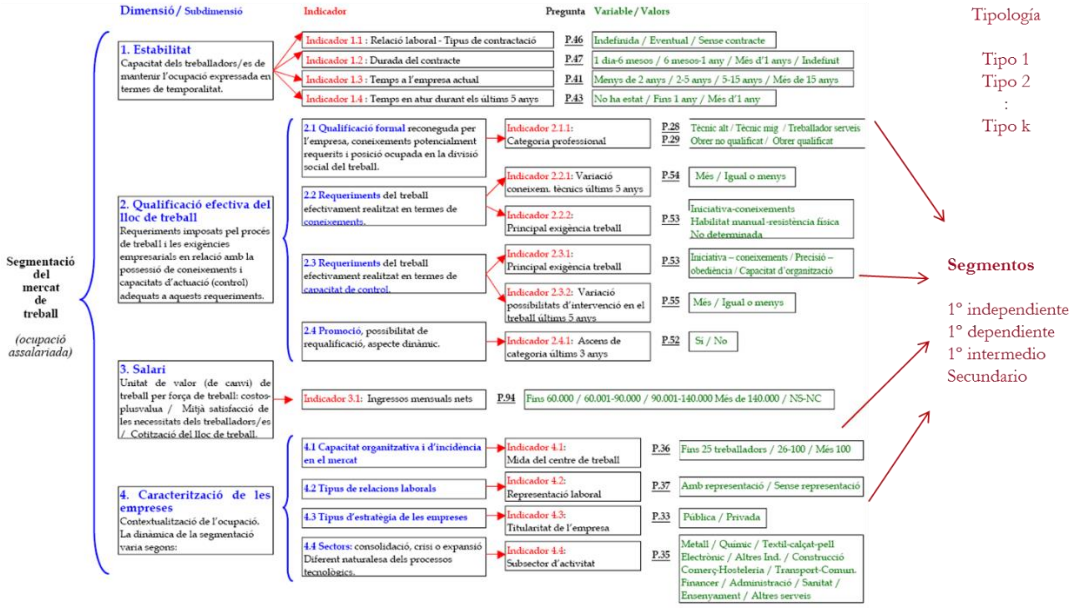


- Clasificación de la población según su relación con la actividad (Encuesta de Población Activa, INE):



3) Tipologías

- Tipos de sociedad
  - F. Tönnies: Comunidad y Asociación
  - É. Durkheim: Solidaridad mecánica y orgánica
- Tipos de autoridad de M. Weber (racional-legal, tradicional, carismática)
- Tipologías de clases sociales y estratificación: Marx, Weber, Erik Olin Wright, J. Goldthorpe, R. Dahrendorf, etc.
- Tipos de capital: tipología de P. Bourdieu (Cultural, Económico, Social, Simbólico)
- Tipos de Estados de Bienestar: tipología de Esping-Andersen (Liberal, Conservador, Socialdemócrata, Mediterráneo-familista)
- Tipos de empleo según la segmentación del mercado de trabajo: institucionalistas y marxistas (sector primario y secundario)



En particular, para la construcción de tipologías se propone una metodología específica que denominamos como estructural y articulada (López-Roldán, 1996a; López-Roldán y Fachelli, 2015).

2. Características, objetivos y modelo de análisis

El objetivo del análisis de clasificación (ACL), genéricamente, consiste en clasificar un conjunto de objetos o individuos *n*, sobre los que se tiene información en términos de *p* variables, en grupos lo más homogéneos. No obstante, la definición más precisa del ACL depende en buena medida de las características matemáticas y técnicas que orientan cada una de las modalidades de clasificación, pues de hecho el ACL es un procedimiento estadístico multivariable que recoge una amplia variedad de procedimientos de clasificación.

El conjunto de estos métodos de clasificación han tenido un desarrollo relativamente reciente en la literatura científica. Los primeros antecedentes se remontan a la psicología ya la obra de Zubin (1938) y Tryon (1939), y en antropología con Driver y

Kroeber (1932). Pero una de las referencias más importantes es la de los biólogos Sokal y Sneath (1963) que establecieron los principios de lo que llamaron la taxonomía numérica. Para las ciencias sociales es notable la referencia de McQuitty (1961), Capecchi (1964, 1966, 1968), Bailey (1983, 1994, 2004), Lorr (1968). A partir de estas contribuciones, dada la importancia de la clasificación en el método científico, y gracias a los rápidos avances de la informática en el tratamiento de la información, el análisis de clasificación experimenta un extraordinario desarrollo que da lugar a la existencia de multiplicidad de métodos y terminologías asociadas.

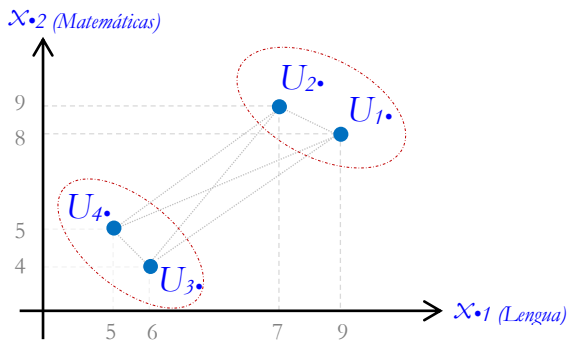
El análisis de clasificación se concibe tanto como un instrumento directo para la constitución de grupos preconcebidos de alguna forma con anterioridad, como un instrumento intermedio de análisis de los datos con un carácter exploratorio. En este sentido permite la construcción de tipologías, o la articulación de hipótesis en la exploración de los datos, o bien la prueba de de hipótesis resultantes de un trabajo estrictamente teórico, permite combinarse con otras técnicas para fin diversas, etc.

El procedimiento general de un análisis de clasificación es bastante simple; a partir de una matriz de datos se trata de agrupar las unidades más similares, más homogéneas. Partimos de una matriz de datos original de n individuos por p variables, o de matrices construidas en base a esta de individuos/ variables, tablas de contingencia, matrices de presencia/ausencia. El ACL, basado en los principios matemáticos de la taxonomía numérica, trata de constituir grupos caracterizados por la densidad de los puntos, por una varianza o dispersión entre ellos, para una dimensión, por una forma y una separación entre los grupos. La mayor parte de los métodos de la ACL comportan simples procedimientos bajo los cuales no hay una gran base de razonamiento estadístico, aunque hay importantes propiedades matemáticas.

Para elaborar una clasificación de unidades similares, homogéneas se plantean dos alternativas: o bien calcular medidas de similitud, como el coeficiente de correlación, o calcular medidas de disimilitud, como las medidas de distancia. Estas medidas, de similitud o disimilitud reciben el nombre de medidas de proximidad.

Para ilustrar estas ideas consideremos el siguiente sencillo ejemplo de cuatro unidades o individuos $U_{i\bullet}$ caracterizados por un espacio de atributos de dos variables (Gráfico III.12.2), por ejemplo: evaluación obtenida en las asignaturas de matemáticas ($x_{\bullet 1}$) y de lengua ($x_{\bullet 2}$), podemos representar esta información en un espacio vectorial de dos dimensiones con cuatro puntos (o vectores) que representan a los individuos.

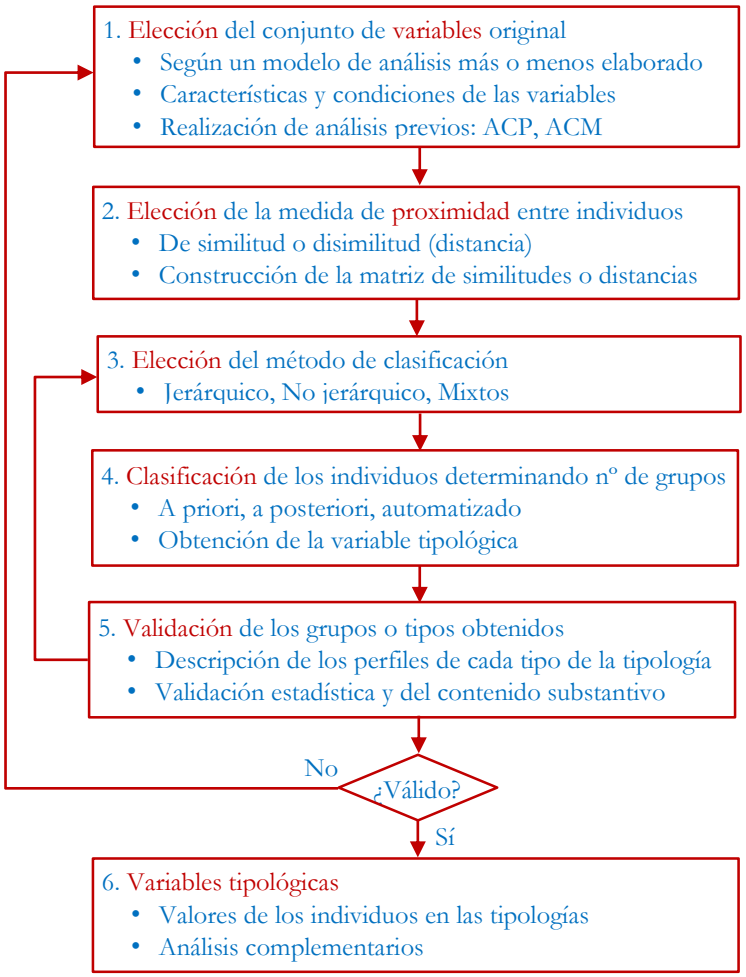
Gráfico III.12.2. Representación de individuos en el espacio



De la observación de esta representación se deduce claramente que los individuos 1 y 2 formarán un grupo o clase (recoge los individuos que sacan unas notas altas en lengua y matemáticas), y los individuos 3 y 4 formarán otro grupo o clase (el de notas no tan altas). Los individuos 1 y 2 están próximos "físicamente" y son próximos "socialmente", tienen una alta similitud, de la misma manera que los individuos 3 y 4. Pero de la misma manera que hemos hecho las comparaciones en términos de similitud podemos hacer las comparaciones en términos de disimilitud: el individuo 1 se distancia mucho del individuo 3 y 4, son disímiles entre sí, al igual que le sucede a la individuo 2, y entre ellos tienen una baja distancia o disimilitud. El mismo comentario se puede hacer en relación de los individuos 3 y 4 en relación a 1 y 2.

Como veremos más adelante considerar las medidas de similitud o de disimilitud, llamadas en conjunto como medidas de proximidad, se corresponde con una decisión y marcan una etapa del proceso de clasificación. Una vez se dispone de la medida que nos permite comparar los casos entre sí se trata de proceder a la formación de los grupos según un procedimiento, un método determinado clasificatorio. Se trata también de escoger este procedimiento, todos ellos caracterizados a partir de formulaciones algorítmicas que establecen la realización de una serie de operaciones recursivas y repetitivas, que dan lugar a la diversidad de métodos de clasificación.

Gráfico III.12.3. Etapas de un análisis de clasificación



El proceso general de un ACL puede estructurarse a partir de las siguientes 6 etapas según se muestra en el Gráfico III.12.3. A continuación pasamos a detallar cada una de estas etapas.

2.1. Etapa 1: Selección de las variables

La aplicación de un análisis de clasificación exige una especial atención en la elección de las variables objeto de estudio, de hecho constituye una de las etapas más críticas en el proceso de investigación, pues de la selección de estas variables depende cualquier conclusión que quiera extraerse del análisis de clasificación. El ACL es un tipo de técnica fundamentalmente exploratoria y descriptiva, donde no se establecen relaciones de dependencia entre las variables, es el conjunto de ellas las que determinan la clasificación de los individuos. La inclusión de unas variables y no de otras, atendiendo a su relevancia en los objetivos del estudio, es crucial para la configuración de los grupos con una composición específica. Por tanto, la consideración de un conjunto determinado de variables deberá estar sujeta siempre a criterios de índole teórica, tanto en la concreción de unos objetivos más o menos precisos e implícitos de una teoría dentro del proceso de investigación como en el contexto más explícito de los postulados de una teoría establecida que guía la clasificación. La tendencia al empiricismo en la aplicación de esta técnica puede conducir a conclusiones precipitadas e incoherentes con la realidad, dada la naturaleza heurística del ACL ante la diversidad de resultados que pueden emerger dada la multiplicidad de circunstancias y criterios que convergen en la obtención de la clasificación apropiada.

Por ello resulta especialmente indicado la realización de otros tipos de análisis previos que permitan la mejor elección de las variables utilizadas, sobre todo si se dispone de un gran número de variables y si estas están correlacionadas entre sí.

Las variables utilizadas en un ACL deben presentar las siguientes características:

- Han respetar en muchos casos la métrica continua, aunque las variables pueden ser cualitativas de tipo binario (dicotómicas, con valores 0/1) o bien pueden ser datos de frecuencias.
- Por otra parte, estas variables deben ser homogéneas y comparables entre sí.
- Es necesario evaluar también si varias variables miden la misma dimensión (si están correlacionadas), y si su importancia es proporcionada, en caso contrario se tiende a reforzar la presencia de alguna característica que condiciona o protagoniza no apropiadamente las características de los grupos resultantes.

Por estos motivos un análisis factorial previo, de componentes principales o de correspondencias, por ejemplo, es un método que nos permite garantizar estas características. Por un lado, como técnica estadística multivariable de reducción de la información, nos proporciona, a partir de las variables originales tratadas, un conjunto nuevo de variables de dimensión significativamente menor mediante la acumulación de la mayor parte de la varianza, y, por otro, al ser variables que forman base, engendran el subespacio vectorial, resultan incorrelacionadas o linealmente independientes. Con las nuevas variables factoriales de un ACP o un ACM podemos obtener las puntuaciones o valores para cada individuo a partir de los cuales proceder a la clasificación. Como el número de componentes que se utilizan es de menor

dimensión que la definida por las variables originales, las distancias entre los puntos o individuos evaluadas a partir de la ACL diferirán de las distancias definidas con las variables originales, pero precisamente en el mejor sentido a efectos del análisis y objetivos de la ACL, pues lo que obtenemos es una nube de puntos donde los individuos se disponen en función de aquellas características que más les discriminan y los hacen diferentes, con las ventajas adicionales mencionados de reducción y incorrelación estadística.

En el momento de seleccionar las variables es necesario también tener en cuenta que las unidades de medida. Los resultados clasificatorios son sensibles al escalamiento, por lo que cuando estas unidades de medida difieren (por ejemplo, una variable es la edad y se mide en años y otro son los ingresos y se mide en euros) es conveniente hacer comparable su métrica estandarizándolas, así se consigue que a la hora de efectuar las comparaciones de las unidades sus diferencias vengan expresadas estrictamente por la medida de similitud empleada y no por el efecto del cambio de unidad de medida al considerar la acción conjunta de varias variables, se consigue así ponderar su importancia relativa, y se evita que afecte a los resultados de la clasificación. La estandarización o tipificación se puede aplicar sobre las variables originales que actúan de criterios clasificatorios. Si utilizamos las variables factoriales obtenidas en un ACP o ACM previo éstas ya vendrán tipificadas.

2.2. Etapa 2: Elección de la medida de proximidad

El criterio de proximidad, bien de similitud o de disimilitud (distancia), es también decisivo en la formación de los grupos o clases. Hay una gran variedad de criterios o medidas que se traducen en índices diversos donde hay que tener en cuenta el nivel de medida de las variables. Las medidas o coeficientes de proximidad se pueden dividir en cuatro tipos según la clasificación de Sneath y Sokal (1973): medidas de distancia, coeficientes de correlación, medidas de asociación para variables binarias y medidas de similitud probabilística.

Nosotros utilizaremos una clasificación basada en el tipo de variable que se considera, según sean: continuas o cuantitativas (intervalo y razón), cualitativas codificadas con dos valores 0 y 1, llamadas binarias o dicotómicas, y finalmente cuando los datos son frecuencias.

2.2.1. *Medidas de proximidad para variables continuas*

Presentamos a continuación una serie de medidas de proximidad para mostrar su variedad, si bien emplearemos fundamentalmente una de ellas, la distancia euclídea.

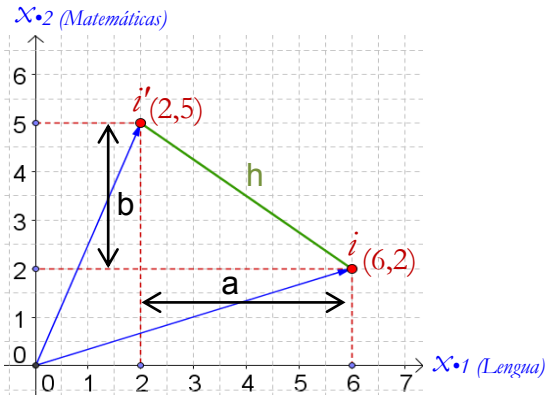
La distancia euclídea o euclidiana es la distancia geométrica entre dos puntos o unidades, que equivale a la longitud de la recta que une ambos puntos en un espacio de p dimensiones. La distancia $d(i,i')$ entre dos individuos i e i' se calcula haciendo la raíz cuadrada de la suma de las diferencias al cuadrado entre los valores de los dos puntos i e i'. Para el caso de dos dimensiones reproduce el conocido Teorema de Pitágoras.

Su expresión es:

$$d(i,i') = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{i'j})^2}$$ *Ecuación 1*

Consideremos el siguiente ejemplo de dos individuos que obtienen las siguientes notas de lengua y matemáticas: 6 y 2 el primero, y 2 y 5 el segundo. La distancia entre ambos se representa y se calcula de la forma siguiente:

Gráfico III.12.4. Distancia entre dos puntos



$$h^2 = a^2 + b^2 = (6-2)^2 + (5-2)^2 = 4^2 + 3^2 = 25 \Rightarrow h = \sqrt{25} = 5$$

También se puede expresar sin la raíz, elevando al cuadrado, se denomina entonces distancia euclídea al cuadrado o cuadrática euclidiana.

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

La fórmula de la distancia euclídea se puede reescribir teniendo en cuenta la correlación de las variables. Cuando existe correlación es conveniente eliminar la redundancia informativa, para ello se utiliza la distancia euclídea en un sistema de coordenadas cartesianas cualquier, con ejes oblicuos:

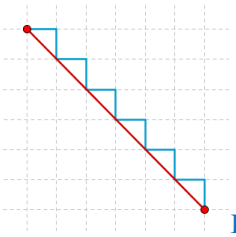
$$d(i,i') = \sqrt{\sum_{j=1}^{p}\sum_{j'=1}^{p} (x_{ij} - x_{i'j}) \cdot (x_{ij'} - x_{i'j'}) \cdot r_{jj'}}$$ *Ecuación 3*

Esta expresión adopta la forma anterior cuando la correlación entre las variables es nula.

Otras medidas de distancias métricas son las siguientes.

- Distancia de Manhattan o bloque (*City-block*) o rectangular. La distancia entre dos puntos es la suma de las diferencias absolutas entre los valores de los dos puntos:

$$d(i,i') = \sum_{k=1}^{p} \left| x_{ij} - x_{i'j} \right| \qquad \textit{Ecuación 4}$$



Distancia de Manhattan y distancia euclídea

- Distancia de Minkowski. Es una distancia más general donde se incluyen los dos casos anteriores. La distancia entre dos elementos es la raíz q-ésima de la suma de las diferencias absolutas a la potencia q-ésima entre los valores para las unidades:

$$d(i,i') = \sqrt[q]{\sum_{j=1}^{p} w_j \cdot \left| x_{ij} - x_{i'j} \right|^q} \qquad \textit{Ecuación 5}$$

donde q es un número real mayor que 0 y donde se contempla la posibilidad de introducir distintos pesos w_j para las variables. Se utiliza con valores q que se inician en 1 (distancia euclídea) o 2 (distancia de Manhattan), con valores que tienden al infinito se convierte en la distancia de Chebychev.

- Distancia de Mahalanobis (Mahalanobis, 1936), llamada también distancia generalizada:

$$\sqrt{d(i,i') = (\vec{E}_{i\bullet} - \vec{E}_{i'\bullet})' \cdot S^{-1} \cdot (\vec{E}_{i\bullet} - \vec{E}_{i'\bullet})} \qquad \textit{Ecuación 6}$$

donde S^{-1} es la inversa de la matriz de varianzas-covarianzas intragrupos de los dos puntos conjuntamente (*pooled within-groups variance-covariance matrix*) que permite la ponderación de las diferencias por las covarianzas. Cuando la correlación entre las variables es cero coincide con la distancia cuadrática euclidiana.

- Coeficiente de correlación. Es el coeficiente de correlación producto momento de Pearson definido en este caso no para variables sino para dos individuos cualesquiera i e i':

$$r_{ii'} = \frac{\sum_{j=1}^{p} (x_{ij} - \bar{x}_{i\bullet})(x_{i'j} - \bar{x}_{i'\bullet})}{\sqrt{\sum_{j=1}^{p} (x_{ij} - \bar{x}_{i\bullet})^2} \cdot \sqrt{\sum_{j=1}^{p} (x_{i'j} - \bar{x}_{i'\bullet})^2}} \qquad \textit{Ecuación 7}$$

donde x_{ij} es el valor de la variable j para el individuo i, y donde $\bar{x}_{i\bullet}$ es la media de los individuos i para los valores de todas las variables. De esta forma se compararían los perfiles propios de cada individuo y se mediría el grado de concordancia o discordancia entre ellos. De igual forma, sin los datos centrados, se obtiene una medida similar que es la distancia que resulta del coseno del ángulo que forman dos vectores.

- Distancia de Chebychev. La distancia entre dos puntos (vectores) es la mayor de sus diferencias a lo largo de cualquier dimensión de coordenadas. Se calcula como diferencia absoluta máxima entre los valores de los puntos:

$$d(i,i') = max\left(\left|x_{ij} - x_{i'j}\right|\right) \qquad \textit{Ecuación 8}$$

para todo j.

2.2.2. *Medidas de proximidad para variables binarias*

Son medidas de similitud o de disimilitud aplicables a variables binarias (dicotómicas) codificadas con 0 y 1, que expresan la ausencia o presencia de una característica. Para ilustrar estas medidas consideraremos el siguiente ejemplo de dos individuos con los valores de 5 variables (por ejemplo si posee o no 5 clases de equipamientos en el hogar) distintos. A partir de esta información se puede construir una tabla de contingencia con cuatro casillas (a, b, c, d) con el cruce de ambos y el recuento de cada pareja de valores:

	$x_{\bullet 1}$	$x_{\bullet 2}$	$x_{\bullet 3}$	$x_{\bullet 4}$	$x_{\bullet 5}$
i	1	1	0	0	1
i'	0	1	0	1	1

i / i'	1	0
1	**a**=2	**b**=1
0	**c**=1	**d**=1

Las diferentes medidas de proximidad tienen en cuenta los valores a, b, c y d. Presentamos seguidamente algunas de estas medidas.

- La medida de similitud de Russell y Rao (1940):

$$\frac{a}{a+b+c+d} = \frac{2}{2+1+1+1} = \frac{2}{5} = 0,4$$

- El coeficiente de concordancia simple o *simple matching* (Sokal y Sneath, 1963):

$$\frac{a+d}{a+b+c+d} = \frac{2+1}{2+1+1+1} = \frac{3}{5} = 0,6$$

- El coeficiente de Roger y Tanimoto (1960) donde las discordancias pesan el doble:

$$\frac{a+d}{a+d+2(b+c)} = \frac{2+1}{2+1+2(1+1)} = \frac{3}{7} = 0,43$$

- Medida de Hamann (1961):

$$\frac{(a+d)-(b+c)}{a+b+c+d} = \frac{(2+1)-(1+1)}{2+1+1+1} = 0,2$$

- El coeficiente de Sokal y Sneath (1963):

$$\frac{a}{a+2(b+c)} = \frac{2}{2+2(1+1)} = \frac{2}{6} = 0,33$$

- El coeficiente de Sokal y Sneath (1963) donde las concordancias pesan el doble:

$$\frac{2(a+d)}{2(a+d)+b+c} = \frac{2(2+1)}{2(2+1)+1+1} = \frac{6}{8} = 0,75$$

- El coeficiente Phi (Sokal y Sneath, 1963):

$$\frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \frac{(2\times1)+(1\times1)}{\sqrt{(2+1)\times(1+1)\times(1+1)\times(1+1)}} = 0,204$$

- El coeficiente de Jaccard o *similarity ratio* (Jaccard, 1908):

$$\frac{a}{a+b+c} = \frac{2}{2+1+1} = \frac{2}{4} = 0,5$$

- El coeficiente de Yule (Yule, 1911):

$$\frac{ad-bc}{ad+bc} = \frac{(2\times1)-(1\times1)}{(2\times1)+(1\times1)} = \frac{1}{3} = 0,33$$

- El coeficiente de Dice (1945):

$$\frac{2a}{2a+b+c} = \frac{2\cdot2}{2\cdot2+1+1} = \frac{4}{6} = 0,67$$

- Distància euclídea binaria:

$$\sqrt{b+c} = \sqrt{1+1} = 1,41$$

- Medida de similitud de Kulczynski (1927):

$$\frac{a}{b+c} = \frac{2}{(1+1)} = 1$$

- Medida de Ochiai (1957):

$$\frac{a}{\sqrt{(a+b)(a+c)}} = \frac{2}{\sqrt{(2+1)(2+1)}} = 0,667$$

2.2.3. *Medidas de proximidad para datos de frecuencias*

La proximidad se puede establecer también entre distribuciones de frecuencias y que expresan perfiles distintos que son comparados, como es el caso de una tabla de contingencia bidimensional al comparar las distintas categorías de una variable entre sí a partir de los datos de que presentan en las otra.

- Medida de chi-cuadrado. Medida basada en la prueba de chi-cuadrado de igualdad para dos conjuntos de frecuencias, con la expresión:

$$d(i,i') = d_{ii'} = \sqrt{\frac{\sum_{j=1}^{p}\left[x_{ij} - E(x_{ij})\right]^2}{E(x_{ij})} + \frac{\sum_{j=1}^{p}\left[x_{i'j} - E(x_{i'j})\right]^2}{E(x_{i'j})}}$$

- Medida de phi-cuadrado. Esta medida es igual a la medida de phi-cuadrado normalizada por la raíz cuadrada de la frecuencia combinada:

$$d(i,i') = d_{ii'} = \sqrt{\frac{1}{n} \cdot \left(\frac{\sum_{j=1}^{p}\left[x_{ij} - E(x_{ij})\right]^2}{E(x_{ij})} + \frac{\sum_{j=1}^{p}\left[x_{i'j} - E(x_{i'j})\right]^2}{E(x_{i'j})} \right)}$$

La utilización de uno u otro índice o medida de proximidad puede dar lugar a resultados de clasificación distintos. La elección del tipo de medida utilizada deberá tomarse siempre en el contexto de la investigación que se lleva a cabo, teniendo en cuenta el nivel de medida de las variables y también los objetivos definidos en el modelo de análisis, si bien en este último caso no hay pautas de elección suficientemente establecidas y son las razones prácticas muchas veces, las que justifican la elección de la medida.

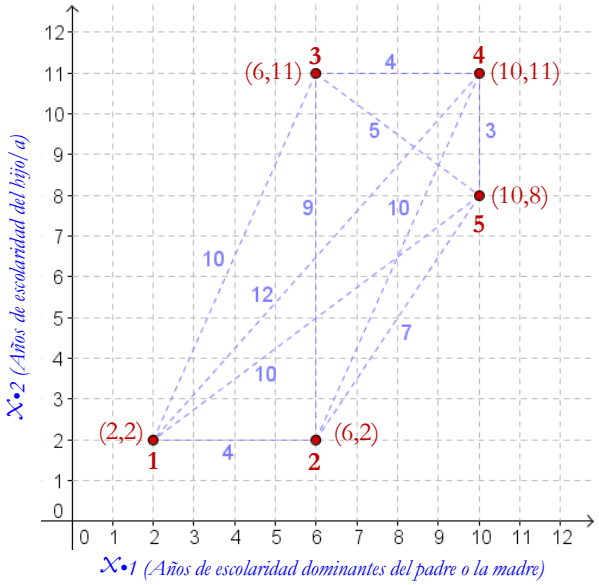
2.2.4. *La matriz de distancias*

Una vez determinados el conjunto de variables pertinentes para la clasificación de los individuos y la medida de proximidad entre ellos, el siguiente paso del proceso de clasificación consiste en la construcción de la matriz de proximidades, de similitud o disimilitud (distancias) entre cada par de unidades. Se trata de una matriz simétrica, matriz de orden $n \times n$, que es el punto de partida de todo el proceso de formación de las clases o de los grupos según el método de construcción elegido, tarea que corresponde a la tercera etapa que presentamos a continuación.

Para ilustrar la creación de una matriz de distancias presentamos el siguiente ejemplo donde hemos considerado un espacio de atributos de 2 dimensiones definido por los años de escolaridad del hijo/a y los años de escolaridad dominantes entre el padre y la madre, donde hemos representado a 5 individuos. Las distancias en todos ellos se calculan a través de la medida euclídea.

El Gráfico III.12.5 representa esta información.

Gráfico III.12.5. Distancias entre cinco individuos



La Tabla III.12.1 por su parte traslada la información de las distancias entre los puntos del gráfico a la forma matricial.

Tabla III.12.1. Matriz de distancias entre 5 individuos

	1	2	3	4	5
1	0	4	10	12	10
2	4	0	9	10	7
3	10	9	0	4	5
4	12	10	4	0	3
5	10	7	5	3	0

Este sencillo ejemplo contrasta con los análisis donde se consideran miles o millones de individuos. Dependiendo del método de clasificación el tratamiento de tantos datos deviene imposible y se exige optar por métodos clasificatorios que admiten el procesamiento de muchos individuos. Lo veremos seguidamente.

2.3. Etapa 3: Elección del método de clasificación

Los diversos métodos de clasificación han aparecido como resultado de los avances realizados en varias disciplinas, los objetos de estudio y las peculiaridades en el tipo de información. De hecho cada método es capaz de generar soluciones notablemente distintas a partir del mismo conjunto de datos analizados, pues cada uno de ellos emplea criterios diferenciados en la constitución de los grupos, lo que representa una fuente de incertidumbre para el investigador/a. Esto nos pone de manifiesto la necesidad, por una parte, de conocer las características propias de cada método para elegir el más apropiado a efectos de la problemática tratada y, por otro lado, considerar el contraste y la validación de los resultados obtenidos por poder determinar finalmente que la clasificación de los grupos obtenida obedece a una lógica no meramente impuesta por el método empleado. A efectos prácticos, no obstante, se

impone muchas veces la disponibilidad de los mismos en el software estadístico utilizado.

Los diferentes métodos de clasificación se pueden dividir en tres grandes tipos: jerárquicos, no jerárquicos (o partición) y mixtos. A continuación presentamos una clasificación de los principales procedimientos clasificatorios y luego detallaremos las características de algunos de ellos.

Tabla III.12.2. Clasificación de los métodos de clasificación

Métodos Jerárquicos	Ascendentes o aglometarivos	Distancias mínimas Distancias máximas Distancia media entre grupos Distancias entre centroides Distancia mediana Ward (mínima pérdida de inercia)
	Descendientes o disociativos	Distancias mínimas Distancias máximas Distancia media entre grupos Distancias entre centroides Distancia mediana Ward (mínima pérdida de inercia) Detector automático de interacción Montéticos Politéticos
Métodos No Jerárquicos o de partición	De reasignación	Centros móviles, K-means Metodo de Forgy Nubes dinámicas con grupos estables Climbing Isodata distancias mínimas Particionamiento alrededor de medoides Mapas autoorganizados
	De búsqueda de densidad	De aproximación tipológica: -Análisis modal de Wishart -Método de Taxmap de Carmichael y Sneath -Método Fortin. De aproximació probabilística: -Método de las combinaciones de Wolf Vecino más cercano
	Directos	Block clustering de Hartigan
	Reductivos	Análisis factorial tipo Q
Mixtos	Clasificación híbrida Clasificación en dos fases	

De todos estos métodos daremos cuenta de algunos de ellos, y en particular profundizaremos en los que pueden tener un mayor interés en el ámbito de las ciencias sociales y de los que a la vez se implementan en software que esté a nuestro alcance, en nuestro caso en SPSS, SPAD y R.

2.3.1. *Métodos jerárquicos*

En el proceso de clasificación por métodos jerárquicos ascendentes o descendentes se emplean algoritmos en los que se procede a la partición de los individuos en agrupaciones o divisiones sucesivas que dan lugar a diferente número de grupos. Las particiones se ordenan en una jerarquía de particiones, la cual se puede representar gráficamente en forma de árbol o dendrograma.

En cada nivel de partición habrá un número diferente de grupos, desde un inicio donde se tienen tantos grupos como unidades hasta llegar a obtener un solo grupo con todas las unidades. El objetivo final consiste en fijar, entre estos dos extremos, el número idóneo de grupos o de clases.

Pero antes de pasar a detallar algunos de los procedimientos de clasificación de tipo jerárquico introduciremos algunos conceptos matemáticos habitualmente implicados en ellos.

El proceso de formación de las clases y de medida de las proximidades se fundamenta matemáticamente a partir de la noción de partición de un conjunto finito, en nuestro caso del conjunto de unidades U. Una partición P_k, con k=1...K, de este conjunto es otro conjunto de partes de U, disjuntas dos a dos, cuya unión es igual a U, es decir, conjunto de partes o grupo de U. A partir de la reunión de todas las particiones posibles de U, $P(U)$, es posible definir un orden parcial de particiones que implican un mayor o menor nivel de agregación de unidades, definiéndose una estructura de redes que muestran todos los posibles caminos, como en el caso de una clasificación jerárquica ascendente, entre una partición que considera al inicio tantas clases como unidades tiene U, P_0, y el nivel máximo de agregación, una partición que considera a todas las unidades formando parte de una misma clase, P_k.

Cada uno de estos posibles caminos se identifica como una cadena de particiones C, es decir, sucesiones de particiones inclusivas que dividen U de forma más o menos desagregada desde P_0 hasta P_k, $C = \{P_0, P_1,..., P_k\}$. Cuando el paso de una partición a otra se produce por la agregación de dos elementos la cadena se denomina binaria.

Una cadena de la red de las partes de U forma un subconjunto totalmente ordenado de $P(U)$, se forma una jerarquía de particiones.

Cuando a cualquier partición o nivel de la jerarquía de las particiones se le puede asignar un valor numérico y se puede, en consecuencia, relacionar con otra partición de orden superior (o inferior) donde aquella se incluye (o la incluye), afirmando que el valor de la primera es inferior (o superior) al de la segunda, entonces se dice además que esta jerarquía es una jerarquía indexada. Por tanto, el hecho de disponer de una jerarquía indexada permite definir una distancia entre las particiones o elementos del conjunto de las partes de U. Este es el principio básico que guía la construcción de las clasificaciones y que se relaciona estrechamente con la noción de distancia en la comparación de los individuos. La distancia entre los valores numéricos de las particiones de una jerarquía indexada es una distancia ultramétrica.

La distancia ultramétrica es un concepto que se expresa de la forma siguiente. En primer lugar, sobre un conjunto de individuos U, se dice que posee una métrica o una distancia, si se cumplen una serie de propiedades matemáticas en U:

a) Condición de separación: $d^2(U_{i\cdot}, U_{i'\cdot}) = 0$ si y solo si $U_{i\cdot} = U_{i'\cdot}$

b) Condición de simetría: $d^2(U_{i\cdot}, U_{i'\cdot}) = d^2(U_{i'\cdot}, U_{i\cdot})$

c) Condición de desigualdad triangular: $d^2(U_{i\cdot}, U_{i'\cdot}) \leq d^2(U_{i\cdot}, U_{i''\cdot}) + d^2(U_{i''\cdot}, U_{i'\cdot})$

Cuando además la distancia cumple una cuarta propiedad más restrictiva:

d) Condición de Krassner: $d^2(U_{i\cdot}, U_{i'\cdot}) \leq \max\{d^2(U_{i\cdot}, U_{i''\cdot}), d^2(U_{i''\cdot}, U_{i'\cdot})\}$

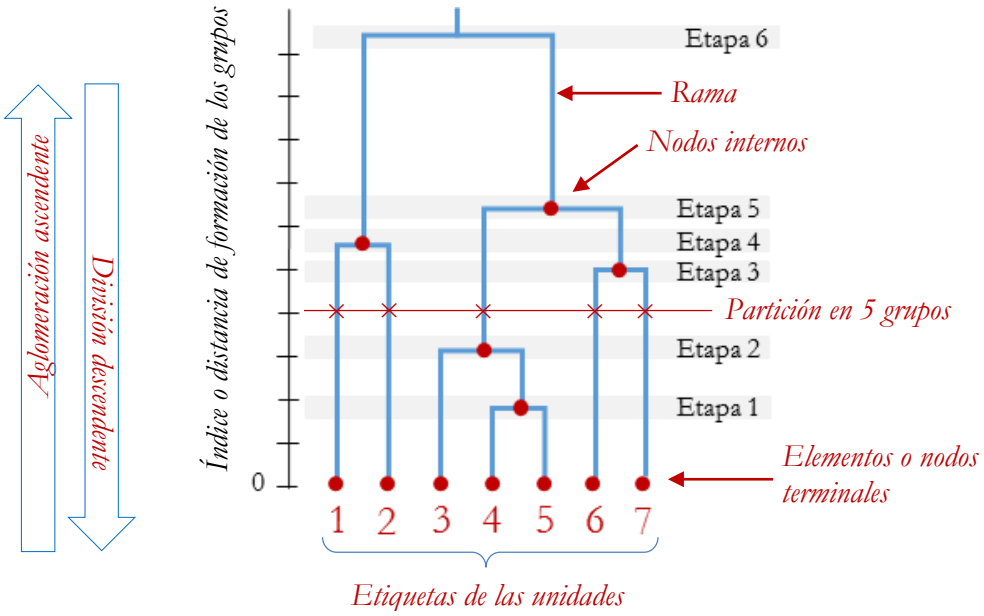
entonces hablamos de distancia ultramétrica que es la medida que nos permitirá comparar, ordenar y diferenciar la jerarquía de particiones. Se demuestra que es equivalente una jerarquía indexada que definir una ultramétrica en el conjunto finito U. Por lo tanto cada una de las cadenas de particiones se le puede asociar una ultramétrica, de forma que d es una función que verifica que $d^2(U_{i\cdot}, U_{i'\cdot})$ es el valor del índice correspondiente a la partición que contiene a la vez $U_{i\cdot}$ y a $U_{i'\cdot}$. La atribución de este índice ultramétrico a cada partición nos informará sobre el nivel al que se formarán los diferentes grupos y la distancia existente entre ellos, información que se utilizará para criterio para decidir el número de grupos.

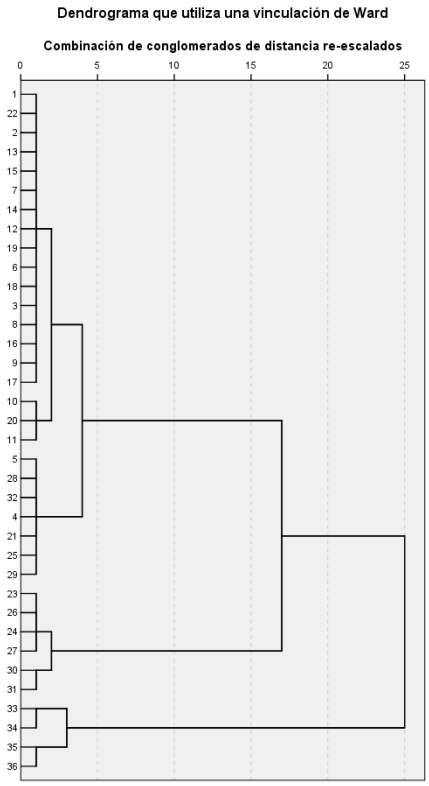
Veamos cómo funcionan estos métodos. El proceso se inicia juntando las dos unidades más cercanas según la información de la matriz de proximidades formando un primer grupo de dos unidades. Si inicialmente teníamos n unidades disponemos de n grupos con una unidad cada uno, y después de la primera etapa donde se unen dos grupos o unidades disponemos de $n-1$ grupos. En la etapa siguiente y progresivamente se trata de reiterar el proceso uniendo de nuevo las dos unidades, que en un momento dado serán unidad y grupo o dos grupos entre sí, más cercanos.

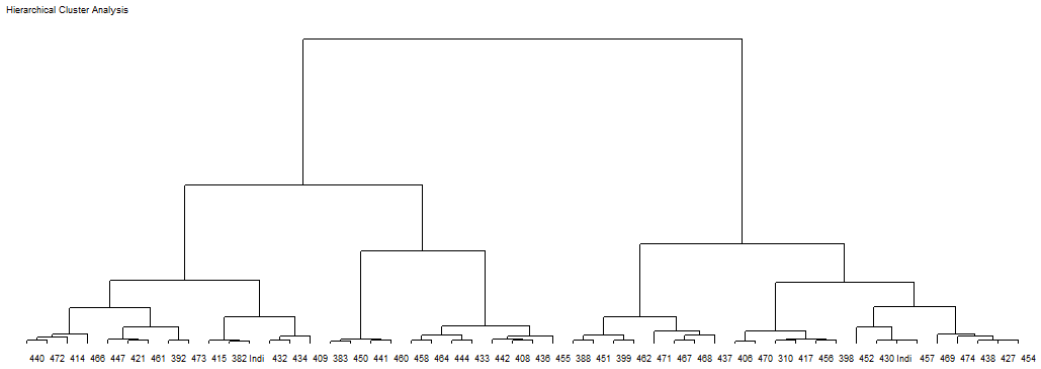
Vemos cómo se realiza esta unión en cualquier etapa. Dada una partición de un conjunto de unidades, por tanto, una clasificación con un número de grupos dado creada en una etapa precedente del proceso, un nueva unidad (que puede ser una unidad individual -un individuo aun no formando parte de ningún grupo- o una unidad grupo -un conjunto de individuos ya agregados en una etapa anterior-) es candidata a unirse con otro unidad, grupo o clase, ya constituida. La medida de la distancia entre esta unidad candidata debe evaluarse según un criterio que determinará cada método jerárquico. En todo caso la respuesta puede ser de inclusión en un grupo o clase, la unión de esta unidad con el grupo o unidad con la que se compara, o de no inclusión o no unión. En cada etapa se evalúan, se comparan, todos los pares posibles, todas las uniones posibles, y de todas ellas sólo se realiza una, la unión que junta las unidades más cercanas según el criterio de cada procedimiento de clasificación. Después de crear esta nueva partición, para poder evaluar de nuevo en la siguiente etapa qué dos unidades o grupos ya creados son los más parecidos se calculará en esta nueva etapa la matriz de proximidades considerando la nueva unidad/grupo agrupada en la etapa anterior. Con esta información se inicia de nuevo el proceso de evaluación.

Este proceso sucesivo se puede representar en un gráfico denominado dendrograma, un gráfico en forma de árbol que ilustra el proceso de agregación y muestra qué grupos se juntan en cada etapa, cuántos grupos hay en cada partición y a qué distancia se forman:

Gráfico III.12.6. Representación de un árbol de clasificación o dendrograma





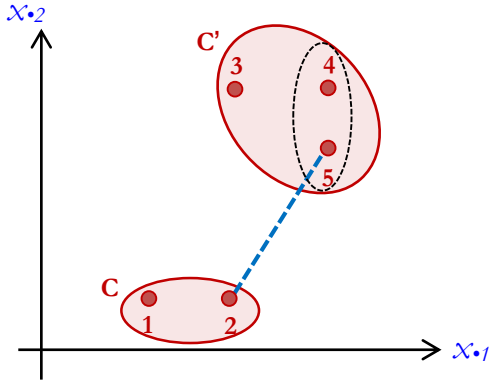


En las clasificaciones jerárquicas con una estrategia divisiva o disociativa el proceso es similar. En este caso no obstante el punto de partida es el conjunto de todas las unidades que se separan en la primera etapa en dos grupos, y se continúa subdividiendo cada uno de ellos en subgrupos con un número menor de unidades, hasta llegar a tener tantos grupos como unidades. De nuevo se trata de decidir entre los dos extremos cuál es la partición, la clasificación más conveniente.

De los comentarios glosados en los párrafos anteriores sobre el proceso de formación de las uniones surge una cuestión relevante. La primera partición se hace a partir de las unidades, entre las que medimos su proximidad optando por una medida (la distancia euclídea, por ejemplo). Es lógico que juntemos, para formar la primera partición, aquellas unidades más próximas. Pero a partir de esta primera partición, una posterior deberá considerar dos unidades en un grupo. La cuestión es ¿cómo se mide la distancia entre una unidad y un grupo formado por dos unidades para poder realizar la correspondiente partición? Y sucesivamente también ¿cómo medir la distancia entre dos grupos? Para dar respuesta existen diversos métodos o criterios, con sus algoritmos, que presentamos a continuación.

2.3.1.1. *Método de distancias mínimas*

El método de distancias mínimas (Sokal y Michener, 1958) se denomina también como vecino más próximo, de salto mínimo, similitud máxima o vinculación simple (*single linkage* o *nearest neighbour*). Para medir la distancia entre una unidad simple y un grupo de unidades, o entre dos grupos de unidades, este método toma la distancia mínima que hay entre dos elementos o unidades. Se trata de comparar la unidad simple con cada una unidad del grupo, o bien, realizar la comparación entre las unidades del primer grupo y las unidades del segundo.



En el ejemplo del gráfico se forma el grupo *C'* uniendo la unidad 3 a los elementos 4 y 5 ya unidos previamente. La distancia de esta nueva clase *C'* con la clase *C* formada por las unidades 1 y 2 resulta de calcular la menor distancia: *d(2,5)*.

Un nuevo elemento se incluye en un grupo, se considera similar, si su distancia es mínima con un solo elemento del grupo, siendo mayor con cualquier otro elemento que no es de este grupo o clase. Siendo *C* y *C'* son dos clases, la distancia entre ambas se define como:

$$d(C,C') = \min\left\{d(U_{i\cdot}, U_{i'\cdot})\right\} \qquad \textit{Ecuación 9}$$

con $U_{i\cdot} \in C$ y $U_{i'\cdot} \in C'$.

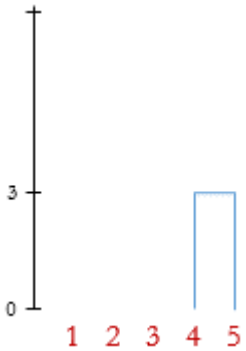
Es un procedimiento muy simple, adecuado para disposiciones de los elementos en cadenas o cuando los grupos están bien separados, pero no cuando hay casos aislados intermedios. Con este procedimiento existe el riesgo de incluir en una misma clase unidades muy distanciadas entre sí.

Por ilustrar en concreto cómo funciona el método de clasificación utilizaremos un sencillo ejemplo donde se consideran 5 individuos, y donde se han evaluado las diferencias entre ellos según una serie de características de acuerdo con una medida de distancia. La matriz de distancias original para establecer la primera partición es la siguiente:

Matriz de distancias 1

	1	2	3	4	5
1	0	4	10	12	10
2	4	0	9	10	7
3	10	9	0	4	5
4	12	10	4	0	3
5	10	7	5	3	0

1) Primera partición. En la primera etapa se agregan los dos individuos más próximos, el [4] y el [5], los que tienen la distancia mínima. La unión de estos dos elementos forma el grupo [4,5]. La formación de este grupo en la primera partición se puede representar en el dendrograma. En la primera etapa del proceso de agregación esta representación adopta la forma siguiente:



La unión se realiza a una distancia de 3 unidades. A continuación, para llevar a cabo la segunda partición, se trata de calcular de nuevo la matriz de distancias con este nuevo elemento o grupo según el criterio de la distancia mínima, para lo que tendremos que conocer la distancia entre el nuevo grupo formado [4,5] y el resto. La distancia entre [4,5] y el resto de individuos se define según el criterio de distancia mínima como la menor de las distancias en relación a todos los elementos. Si comparamos [4,5] con [1], la distancia es:

$$d([4,5],[1]) = \text{mín}\{d([4],[1]), d([5],[1])\} = \text{mín}\{12,10\} = 10$$

Si comparamos [4,5] con [2] y [3] obtenemos los siguientes resultados:

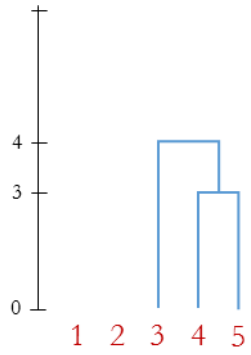
$$d([4,5],[2]) = \text{mín}\{d([4],[2]), d([5],[2])\} = \text{mín}\{10,7\} = 7$$
$$d([4,5],[3]) = \text{mín}\{d([4],[3]), d([5],[3])\} = \text{mín}\{4,5\} = 4$$

Con estos cálculos ya podemos formar la nueva matriz de distancias que ahora contempla tan solo cuatro grupos:

Matriz de distancias 2

	1	**2**	**3**	**[4,5]**
1	0			
2	4	0		
3	10	9	0	
[4,5]	10	7	4	0

2) Segunda partición. Ahora planteamos la siguiente agrupación. La menor distancia entre los individuos/grupos es 4. Entre las dos opciones de la tabla se elige la unión entre el [3] y el [4,5], por tanto, se unen para formar el grupo [3,4,5]. El dendrograma se completa con esta unión de la forma siguiente:



La unión se produce a una distancia de 4. Con el nuevo grupo formado calculamos la distancia de [3,4,5] en relación al resto de elementos, [1] y [2]:

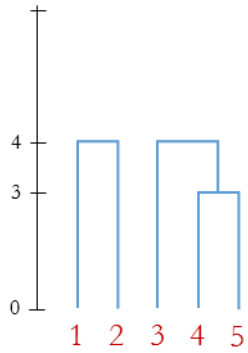
$$d([3,4,5],[1]) = \text{mín}\{d([3],[1]), d([4,5],[1])\} = \text{mín}\{10,10\} = 10$$
$$d([3,4,5],[2]) = \text{mín}\{d([3],[2]), d([4,5],[2])\} = \text{mín}\{9,7\} = 7$$

La nueva matriz de distancias es:

Matriz de distancias 3

	1	2	[3,4,5]
1	0		
2	4	0	
[3,4,5]	10	7	0

3) Tercera partición. Iniciamos ahora la tercera etapa. La menor distancia entre los individuos/grupos es 4 y se da entre [1] y [2], por tanto, se unen para formar el grupo [1,2], también a una distancia de 4. El dendrograma queda:



La distancia entre [1,2] y [3,4,5], en relación a D es:

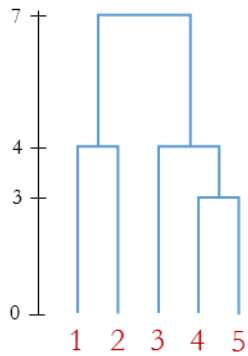
$$d([1,2],[3,4,5]) = \min\{d([1],[3,4,5]), d([2],[3,4,5])\} = \min\{10,7\} = 7$$

La nueva y última matriz de distancias es:

Matriz de distancias 4

	1	[3,4,5]
[1,2]	0	
[3,4,5]	7	0

4) Cuarta partición. En la cuarta y última etapa se agregan los últimos dos grupos, [1,2] y [3,4,5] para forma el grupo con todas las unidades [1,2,3,4,5] a una distancia de 7. El dendrograma queda finalmente como:

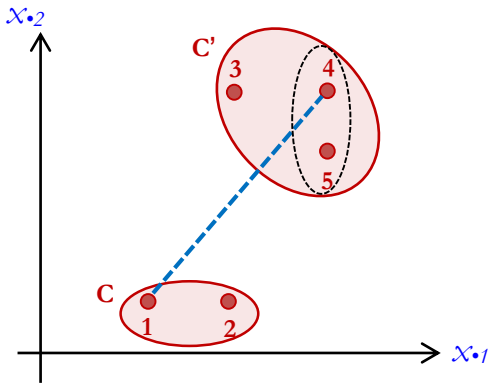


2.3.1.2. *Método de distancias máximas*

El método de distancias máximas (Sorensen, 1948), también llamado del vecino más lejano, del diámetro, del salto máximo, de similitud mínima o vinculación completa (*complete linkage* o *farthest neighbour*), toma como distancia entre una unidad y un grupo o bien entre dos grupos, la distancia máxima que resulta de la comparación de cada elemento del primer grupo y cada uno del segundo. Evalúa la distancia entre los dos puntos o grupos más lejanos en cada partición. Un individuo es asignado a un grupo determinado si está más próximo de todos los individuos de este grupo que de todos los individuos de otro grupo, es decir, cuando la distancia máxima es mínima en relación a la distancia con cualquier otro elemento de fuera. Con este método se forman grupos más compactos, pero no es adecuado ante grupos internamente heterogéneos, aunque estén bien separados, impone la condición de que los puntos más lejanos estén próximos, pero no se ve condicionado por la existencia de elementos intermedios. Siendo C y C' dos clases, la distancia entre ambas se define como:

$$d(C,C') = \max\left\{d(U_{i\bullet}, U_{i'\bullet})\right\} \qquad \textit{Ecuación 10}$$

con $U_{i\bullet} \in C$ y $U_{i'\bullet} \in C'$.



En el ejemplo del gráfico se forma el grupo *C'* uniendo la unidad 3 a los elementos 4 y 5 ya unidos previamente. La distancia de esta nueva clase *C'* con la clase *C* formada por las unidades 1 y 2 resulta de calcular la mayor distancia: *d(1,4)*.

Consideremos el mismo ejemplo anterior. El proceso de agrupación que ahora comentamos crea sucesivamente estas cuatros matrices de distancia:

Matriz de distancias 1

	1	2	3	4	5
1	0				
2	4	0			
3	10	9	0		
4	12	10	4	0	
5	10	7	5	3	0

Matriz de distancias 2

	1	2	3	[4,5]
1	0			
2	4	0		
3	10	9	0	
[4,5]	12	10	5	0

Matriz de distancias 3

	[1,2]	3	[4,5]
[1,2]	0		
3	10	0	
[4,5]	10	5	0

Matriz de distancias 4

	1	[3,4,5]
[1,2]	0	
[3,4,5]	10	0

En la primera etapa se agregan igualmente los dos individuos más próximos, el [4] y el [5], los que tienen la distancia más pequeña. La unión de estos dos elementos forma el grupo [4,5] y se obtiene la primera partición. A continuación se trata de calcular de

nuevo la matriz de distancias con este nuevo elemento o grupo según el criterio de la distancia máxima:

$$d([4,5],[1]) = \text{máx}\{d([4],[1]), d([5],[1])\} = \text{máx}\{12,10\} = 12$$
$$d([4,5],[2]) = \text{máx}\{d([4],[2]), d([5],[2])\} = \text{máx}\{10,7\} = 10$$
$$d([4,5],[3]) = \text{máx}\{d([4],[3]), d([5],[3])\} = \text{máx}\{4,5\} = 5$$

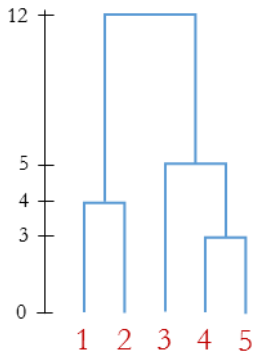
Con estos cálculos construimos la nueva matriz de distancias 2 e iniciamos la segunda etapa. La segunda partición surge a partir de unir los individuos más próximos: el [1] y el [2], por tanto, se unen a la distancia de 4 para formar el grupo [1,2]. La distancia de [1,2] en relación a [3] y [3,4] es:

$$d([1,2],[3]) = \text{máx}\{d([1],[3]), d([2],[3])\} = \text{máx}\{4,10\} = 10$$
$$d([1,2],[4,5]) = \text{máx}\{d([1],[4,5]), d([2],[4,5])\} = \text{máx}\{12,10\} = 12$$

Con estos datos construimos de nuevo la nueva matriz de distancias y pasamos a la tercera etapa. En la tercera partición los individuos o grupos más próximos son [3] y el [4,5], por lo tanto, se unen para formar el grupo [3,4,5]. La distancia de [3,4,5] en relación [1,2] es:

$$d([1,2],[3,4,5]) = \text{máx}\{d([1,2],[3]), d([1,2],[4,5])\} = \text{máx}\{10,10\} = 10$$

Así obtenemos la última matriz de distancias. La última etapa supone la unión de todos los grupos para formar uno único grupo con los cinco individuos. El proceso de agregación que acabamos de describir varía en relación al método anterior de distancias mínimas y su dendrograma tiene la siguiente representación:



2.3.1.3. *Método de la distancia media entre grupos*

Veremos seguidamente cuatro procedimientos que tratan situarse entre las características de los dos métodos anteriores al basarse en la estimación de la "distancia media" entre las unidades de las dos clases o grupos. En cada caso la noción de media se operativizará de forma distinta.

Consideramos en primer lugar el método de la distancia media entre grupos (Sokal y Michener, 1958) que se conoce también como vinculación promedio inter-grupos

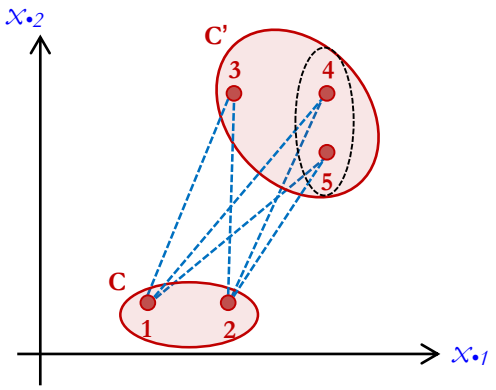
(*average linkage between groups*), UPGMA (*Unweighted Pair-Group Method using atithmetic averages*) o de similitud media.

Dos grupos se unirán cuando esta distancia media, no ponderada, que resulta de considerar todas las distancias entre cada par de puntos que pertenecen a los dos grupos candidatos a unirse, sea la más pequeña después de comparar todos los pares. Se define como:

$$d(C,C') = \frac{\sum_{i}\sum_{i'} d(U_{i\bullet},U_{i'\bullet})}{Card(C)\times Card(C')}$$

Ecuación 11

con $U_{i\bullet} \in C$ y $U_{i'\bullet} \in C'$ y donde *Card(C)* es el número de unidades del primer grupo o clase y *Card(C')* es el número de unidades del segundo.



En el ejemplo del gráfico se forma el grupo *C'* uniendo la unidad 3 a los elementos 4 y 5 ya unidos previamente. La distancia de esta nueva clase *C'* con la clase *C* formada por las unidades 1 y 2 resulta de calcular la suma de todas las distancias: *d(1,3)*, *d(1,4)*, *d(1,5)*, *d(2,3)*, *d(2,4)* y *d(2,5)*, dividiendo por 6, el número de distancias.

El método se califica de no ponderado (*unweighted*) porque no tiene en cuenta las unidades de cada grupo para medir las distancias entre ellos. El denominador de la fórmula anterior *Card(C)×Card(C')* sólo expresa el número de combinaciones entre un grupo y otro para calcular la media de las distancias, no es una ponderación.

Con los datos del ejemplo que seguimos, las matrices de distancias del proceso de clasificación jerárquico ascendente son las siguientes:

Matriz de distancias 1

	1	2	3	4	5
1	0				
2	4	0			
3	10	9	0		
4	12	10	4	0	
5	10	7	5	3	0

Matriz de distancias 2

	1	2	3	[4,5]
1	0			
2	4	0		
3	10	9	0	
[4,5]	11	8,5	4,5	0

Matriz de distancias 3

	[1,2]	3	[4,5]
[1,2]	0		
3	9,5	0	
[4,5]	8,875	5	0

Matriz de distancias 4

	1	[3,4,5]
[1,2]	0	
[3,4,5]	2,396	0

En la primera etapa se agregan de nuevo los dos individuos más próximos, el [4] y el [5] para formar el grupo [4,5]. Las nuevas distancias se obtienen sumando las distancias entre todos los elementos de las unidades/grupos y dividiendo por el producto del número de elementos:

$$d([4,5],[1]) = \frac{d([4],[1]) + d([5],[1])}{2 \times 1} = \frac{12+10}{2} = 11$$
$$d([4,5],[2]) = \frac{d([4],[2]) + d([5],[2])}{2 \times 1} = \frac{10+7}{2} = 8,5$$
$$d([4,5],[3]) = \frac{d([4],[3]) + d([5],[3])}{2 \times 1} = \frac{4+5}{2} = 4,5$$

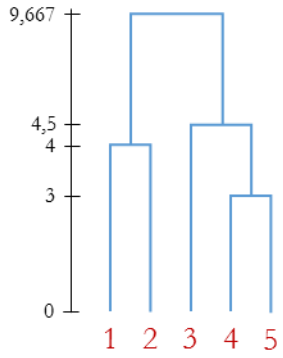
Se unen [1] y [2] y las nuevas distancias calculadas son:

$$d([1,2],[3]) = \frac{d([1],[3]) + d([2],[3])}{2 \times 1} = \frac{10+9}{2} = 9,5$$
$$d([1,2],[4,5]) = \frac{d([1],[4,5]) + d([2],[4,5])}{2 \times 2} = \frac{11+8,5}{4} = 4,875$$

Se unen [3] y [4,5] y la última distancia es:

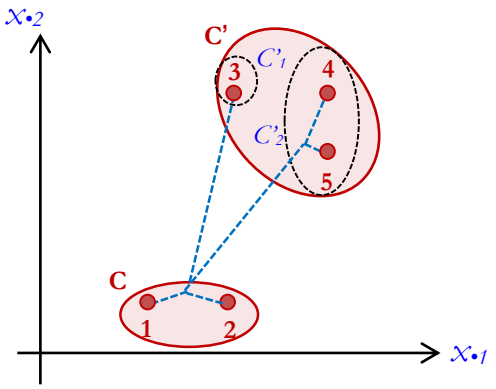
$$d([1,2],[3,4,5]) = \frac{d([1,2],[3]) + d([1,2],[4,5])}{2 \times 3} = \frac{9,5+4,875}{6} = 2,396$$

El árbol de agregación en este caso es el mismo que se obtiene con el procedimiento de distancias mínimas, aunque a con coeficiente de formación de los grupos diferentes.



2.3.1.4. *Método de la distancia media intra grupos*

El método de la distancia media intra-grupos (McQuitty, 1966) o de vinculación promedio intra-grupos (*average linkage within groups*) o WPGMA (*Weighted Pair-Group Method using arithmetic averages*) o de similitud media ponderada, es una variante del método anterior que supone la ponderación del cálculo promedio. Si consideramos una unidad (o un grupo) C'_1 que se une a otro ya formado C'_2 para formar la clase C', la distancia de C' respecto a la clase C se define como el promedio ponderado de las distancias de cada uno de los dos subgrupos C'_1 y C'_2 a C.



En el ejemplo del gráfico anterior se forma el grupo *C'* uniendo la unidad 3 a los elementos 4 y 5 ya unidos previamente. La distancia de esta nueva clase *C'* con la clase *C* formada por las unidades 1 y 2 resulta de calcular la media entre las distancias *d([1,2],[3])* y *d([1,2],[4,5])*.

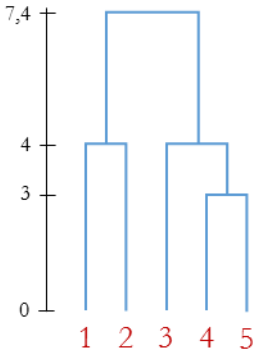
En general, con $n_{C'1}$ unidades de C'_1, $n_{C'2}$ unidades de C'_2 y n_C de C, su expresión es:

$$d(C,C') = \frac{\sum_{i'=1}^{n_{C'1}+n_{C'2}} \sum_{i=1}^{n_C} d(U_{i\bullet},U_{i'\bullet})}{(n_{C'1}+n_{C'2})\cdot n_C} = \frac{n_{C'1}\cdot d(C'_1,C)+n_{C'2}\cdot d(C'_2,C)}{n_{C'1}+n_{C'2}}$$

Ecuación 12

con $U_{i\bullet} \in C$ y $U_{i'\bullet} \in C'$

El análisis de los datos del ejemplo no arroja los mismos resultados de agrupación y varía al final el coeficiente que mide la distancia a la que se forma los grupos.



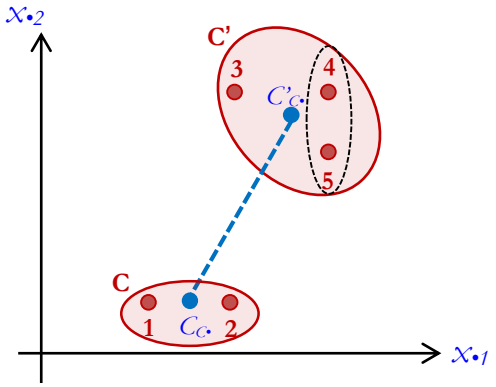
2.3.1.5. *Método de las distancias medianas*

El método de las distancias medianas (Edwards y Cavalli-Sforza, 1965) es también conocido como de vinculación de medianas (*median*) o UPGMC (*Unweighted Pair-Group Method using centroids*) y se basa en la medida de distancia entre centroides (las medias de las variables para los individuos de las clases), en este caso con los datos no ponderados.

Se define la distancia entre una unidad (o grupo) y un grupo, como la distancia de la unidad (que es su propio centroide o punto que representa la media aritmética) y el

centroide o media aritmética de todas las unidades del grupo teniendo en cuenta todas las variables que las caracterizan (o entre el centroide de un grupo, o media aritmética que lo representa, y el centroide del segundo grupo). Se llama *unweighted* (sin ponderación) porque no se tienen en cuenta el número de casos de cada uno de los grupos, el peso asignado a cada grupo es el mismo.

Se unirán así los grupos que tengan una menor distancia consiguiendo que grupos pequeños tengan un efecto igual que los grupos grandes a la hora caracterizar las clases. Este procedimiento exige igualmente la utilización de la distancia euclídea.



En el gráfico se forma el grupo C' uniendo la unidad 3 a los elementos 4 y 5 ya unidos previamente, siendo la distancia de esta nueva clase C' con la clase C la que resulta entre los centros de cada grupo $C_{C\bullet}$ y $C'_{C\bullet}$.

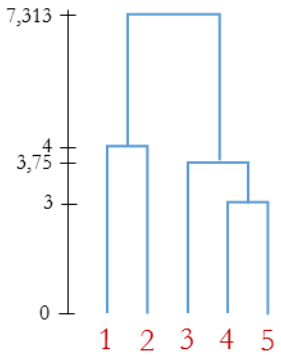
La distancia se puede expresar como:

$$d(C,C') = d(C_{c\bullet}, C'_{c\bullet}) \qquad \textit{Ecuación 13}$$

donde $C_{c\bullet}$ y $C'_{c\bullet}$ son las medias aritméticas o puntos centroides de cada grupo:

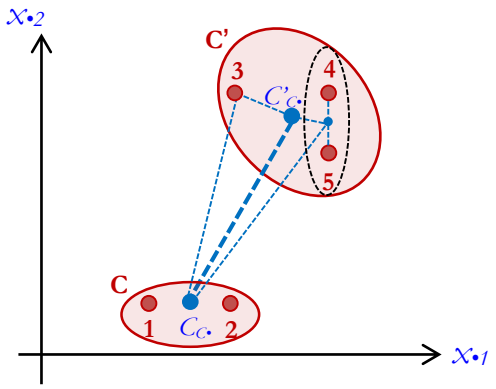
$$C_{c\bullet} = \frac{\sum_{i=1}^{n_c} I_{i\bullet}}{n_c} \quad \text{y} \quad C_{c'\bullet} = \frac{\sum_{i'=1}^{n_{c'}} I_{i'\bullet}}{n_{c'}}$$

El proceso de agregación aplicando este método a los datos del ejemplo da lugar al siguiente dendrograma:



2.3.1.6. *Método de las distancias entre centroides*

El método de las distancias entre centroides (Gower, 1967) o vinculación de centroides (*centroid*), o WPGMC (*Weighted Pair-Group Method using centroids*), sigue el mismo criterio que en el caso anterior, es decir, calcula la distancia entre las medias de los dos grupos para todas las variables, pero ponderando por el tamaño de las clases.



En este ejemplo gráfico la distancia entre C y C' (donde se unen 3 junto con el grupo de 4 y 5 previamente formado) es la mediana del triángulo formado por el centroide del grupo *C*, el centroide de la unidad [3] y el centroide del grupo [4,5].

Considerando la distancia cuadrática euclidiana, con $n_{C'1}$ unidades de C'_1, $n_{C'2}$ unidades de C'_2 y n_C de C, la fórmula de cálculo de la distancia se puede expresar de la forma siguiente:

$$d^2(C,C') = \frac{n_{C'1}}{n_{C'1}+n_{C'2}} d(C'_1,C) + \frac{n_{C'2}}{n_{C'1}+n_{C'2}} d(C'_2,C) - \frac{n_{C'1}\cdot n_{C'2}}{n_{C'1}+n_{C'2}} d(C'_1,C'_2) \quad \textit{Ecuación 14}$$

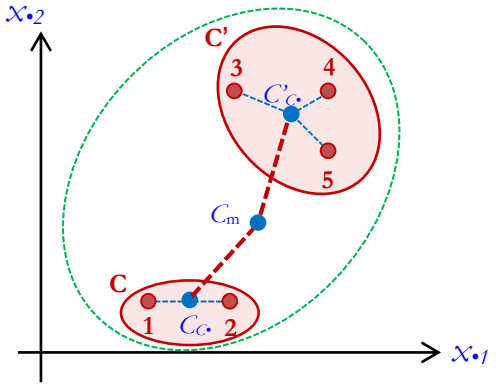
Los resultados que se obtienen con este método son los mismos que con lo anterior, el dendrograma y las distancias son las mismas.

2.3.1.7. *El método Ward*

El método de Ward (1963) o de vinculación de ward, también denominado de mínima pérdida de inercia, de mínima varianza, de mínimo error cuadrático o de agregación según la varianza. Difiere de los vistos hasta ahora por tener en cuenta un criterio de variabilidad.

El método ward, de extendido uso en ciencias sociales, consiste en un proceso progresivo de agregación de las unidades/grupos de manera que en cada etapa se unan aquellos dos elementos que supongan la mínima pérdida de inercia (o varianza), es decir, que junte los dos grupos que hagan disminuir menos la varianza entre los grupos. Por tanto, es un criterio de optimización de la varianza explicada por la unión y

minimización de la varianza residual. Se trata de un procedimiento que exige la utilización de la distancia cuadrática euclidiana.



Si consideramos un número de casos n en el espacio euclidiano definido por p variables o dimensiones, cada punto o individuo $U_{i\bullet}$ (vector de p componentes) en este espacio tiene una masa o peso relativo m_i, que aquí consideraremos de valor unidad, por tanto, siendo la suma de todos los puntos: $\sum_{i=1}^{n} m_i = M = n$, es decir, el número de casos.

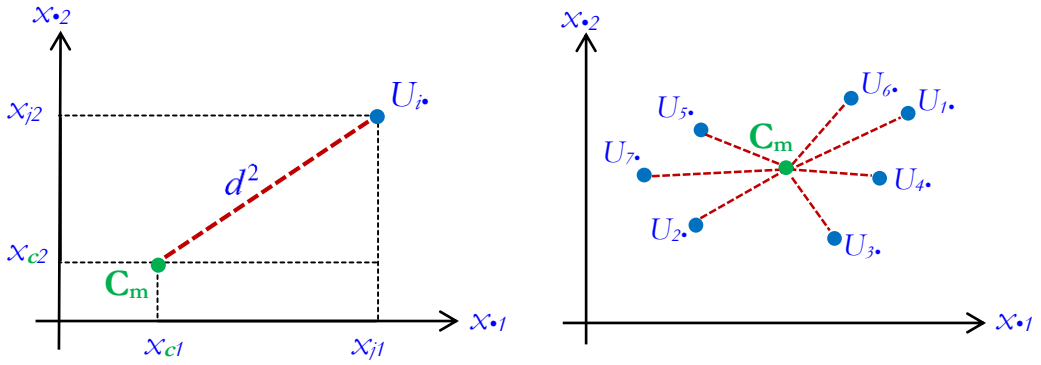
Por otro lado en esta nube de puntos existe un centro de masas o centro de gravedad que se define como la media global del conjunto de puntos o centro de la nube y que se obtiene calculando la media de las p variables para todo los puntos, es decir, es el vector $C_m = (\bar{x}_{\bullet 1}, \bar{x}_{\bullet 2}, \cdots, \bar{x}_{\bullet p})$ donde:

$$\bar{x}_{\bullet j} = \frac{\sum_{i=1}^{n} m_i \cdot x_{ij}}{n} \quad \text{con } m_i = 1 \text{ para todo } i.$$

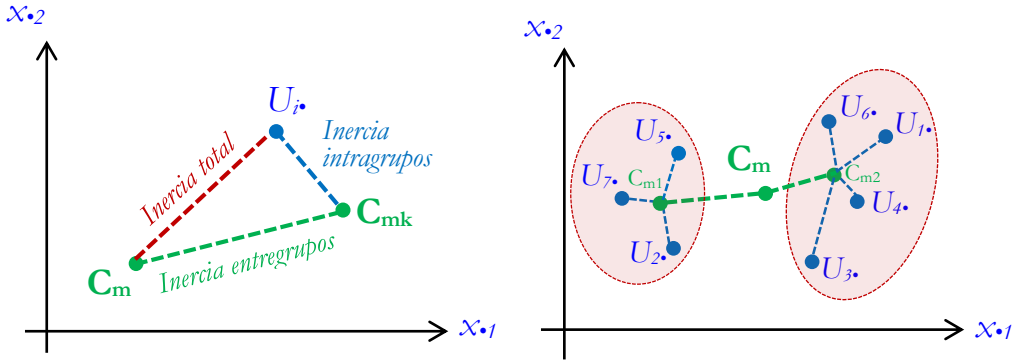
La inercia total de la nube de puntos en relación al centro de masas, I_{Cm}, se define como una medida de dispersión o de variabilidad de la misma expresada a través de la suma de las distancias euclídeas de cada punto $U_{i\bullet}$ al centro de masas global, C_m:

$$I_{Cm} = \sum_{i=1}^{n} m_i \cdot d^2(U_{i\bullet}, C_m) = \sum_{i=1}^{n}\sum_{j=1}^{p} m_i \cdot (x_{ij} - \bar{x}_{cj})^2 \qquad \textit{Ecuación 15}$$

Gráfico III.12.1. Representación de la inercia y su descomposición. Distancia de un punto en el centro de masas y al centro de su grupo.



(a) Inercia de un punto *(b) Inercia de un conjunto*



(c) Descomposición de la inercia de un punto

(d) Descomposición de la inercia de un conjunto según la relación de Huygens

En cada grupo posible k de una partición también existe un centro de gravedad propio del conjunto de individuos que lo forman, C_{mk}, y por tanto una inercia en relación al centro de masas de este grupo, I_{Cmk}. Se puede demostrar que la inercia total será descomponible en suma de inercias según la relación de Huygens:

$$Inercia\ Total = Inercia\ intragrupos + Inercia\ entregrupos$$

$$I_T = I_I + I_E$$

a) Inercia intragrupos: la inercia interna de cada grupo que se calcula sumando las distancias al cuadrado entre los n_k puntos U_i• del grupo k al centro de masas (o centroide o punto de las medias aritméticas), C_{mk}, de este grupo.

b) Inercia entregrupos: la inercia externa de los grupos que se calcula sumando las distancias al cuadrado del centro de masas de cada grupo (o centroide o punto de las medias aritméticas), C_{mk}, al centro global de la nube, C_m, ponderando cada distancia por el número de unidades n_k del grupo.

Las inercias adoptan la expresión:

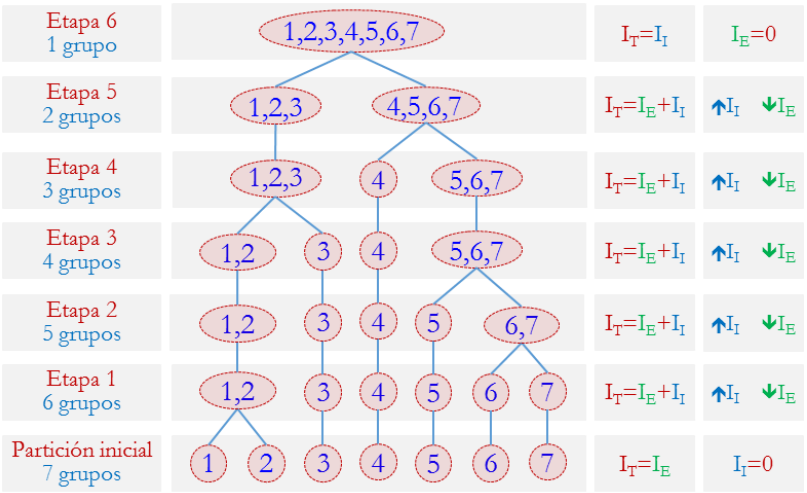
$$I_{Cm} = \sum_{k=1}^{K}\left(\sum_{i=1}^{n_k} m_i \cdot d^2(E_{i\bullet}, C_{mk})\right) + \sum_{k=1}^{K} m_k \cdot d^2(C_{mk}, C_m) \qquad \textit{Ecuación 16}$$

La aplicación del método jerárquico ascendente Ward implica la agregación sucesiva de pares de unidades en un proceso iterativo donde se irá evaluando cada unión posible de unidades o clases en términos de la variación de estas inercias. En un inicio donde se tienen tantos grupos como unidades, la inercia total, que permanece siempre constante, será igual a la inercia entregrupos mientras que la inercia intragrupos será igual a cero, pues aún no se ha formado ninguna agrupación e internamente cada grupo es de una unidad y su inercia o varianza es nula. Como resultado de la aglomeración de las unidades en grupos según las distintas particiones esta inercia total se irá distribuyendo sucesivamente entre la inercia entregrupos y la inercia intragrupos de tal forma que medida que disminuye el número de grupos aumenta la inercia intra disminuyendo la entre y así hasta la última partición donde sólo existirá un grupo y

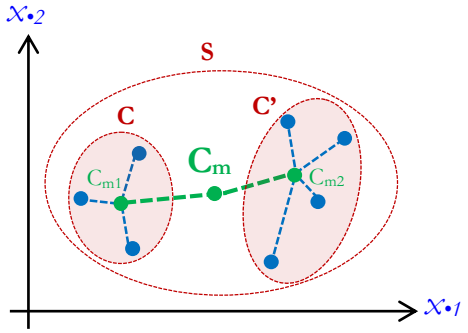
donde, por tanto, la inercia total será inercia intra mientras que la inercia entre será cero:

$$I_{Total} = constante \Rightarrow \triangle I_{Total} = 0$$
$$\triangledown I_{Entre} = \triangle I_{Intra}$$

Podemos utilizar la representación del árbol de agregación para expresar esta idea:



El criterio de agregación de dos unidades en la primera etapa descansa sobre el principio de la mínima pérdida de inercia, es decir, se trata de unir en un mismo grupo en cada nueva partición aquellos dos casos de todos los pares posibles cuya agregación suponga un menor incremento de inercia intra, pues esto nos indicará que la variabilidad interna de este grupo será la menor posible y, por tanto, que serán los dos casos más homogéneos. En cada nueva partición tendremos un aumento de la inercia intragrupos, de variabilidad interna, pero que será el menor incremento posible, y que se produce a costa siempre de la inercia entregrupos. De ello se puede deducir, a efectos de la elección de una partición, que aquella que suponga un incremento o un salto de inercia entre importante es como resultado de la unión de dos grupos cuya agregación implica una alta heterogeneidad, en consecuencia, antes de considerarlos como un solo grupo es preferible optar por mantenerlos separados y elegir la partición anterior.



Así pues, la variación de la inercia, resultante de la unión de dos grupos para formar cada nueva partición, vendrá expresada por los cambios en la distribución de las inercias intra y entre. Cuando se unen dos puntos o grupos cualesquiera *C* y *C'*, disjuntos (véase gráfico adjunto) para formar el grupo *S*, la inercia total de ambos en

relación al centro de masas es igual a la suma de la inercia respecto al propio centro de masas, inercias intra I_{Cm1} e I_{Cm2}, más la inercia de cada grupo (los puntos centro de masas de C y C', C_{m1} y C_{m2}, con m_1 y m_2 como valor de masa) en relación al centro de masas global, C_m:

$$I_{Cm}(A \cup B) = \underbrace{I_{Cm1} + I_{Cm2}}_{Inercia\ Intra} + \underbrace{m_1 \cdot d^2(C_{m1}, C_m) + m_2 \cdot d^2(C_{m2}, C_m)}_{Inercia\ Entre} \qquad Ecuación\ 17$$

Desarrollando el término de la inercia entre a partir de la sustitución del centro de masas por su valor:

$$C_m = \frac{m_1 \cdot C_{m1} + m_1 \cdot C_{m1}}{m_1 + m_2}$$

se obtiene que la inercia total es:

$$I_{Cm}(A \cup B) = \underbrace{I_{Cm1} + I_{Cm2}}_{Inercia\ Intra} + \underbrace{\frac{m_1 \cdot m_2}{m_1 + m_2} \cdot d^2(C_{m1}, C_{m2})}_{Inercia\ Entre} \qquad Ecuación\ 18$$

Si ahora consideramos los dos grupos como elementos de uno solo, S, la inercia total permanecerá invariable y, por tanto: $I_{Cm}(A \cup B) = I_{Cm}(S)$. Pero en S, como tenemos un solo grupo, la inercia entregrupos es nula, todo es inercia intragrupos. En consecuencia, como la inercia total es constante, la variación de inercia como resultado de la formación del grupo S será igual a la variación de inercia intra:

Si $I_{Cm}(A \cup B) = Inercia\ intragrupos(A \cup B) + Inercia\ entregrupos(A \cup B)$

y $I_{Cm}(S) = Inercia\ intragrupos(S) = I_{Cm}(A \cup B)$, por tanto,

$I_{Cm}(S) = Inercia\ intragrupos(A \cup B) + Inercia\ entregrupos(A \cup B)$, es decir:

$I_{Cm}(S) = Inercia\ intragrupos(A \cup B)$

y, $Variación\ de\ Inercia\ intragrupos = Inercia\ entregrupos(A \cup B)$

es decir,

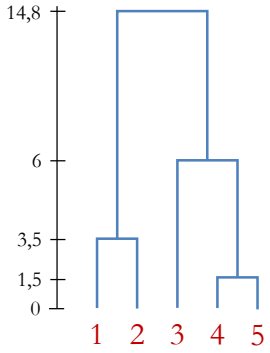
$$\frac{m_1 \cdot m_2}{m_1 + m_2} \cdot d^2(C_{m1}, C_{m1}) = \delta(P_p, P_{p+1}) \qquad Ecuación\ 19$$

Esta variación de inercia debe ser la menor posible y permite ser interpretada como un nuevo índice δ que mide la distancia a la que se forma la nueva partición P_{p+1} al pasar de k grupos, en la partición anterior P_p, a $k-1$ grupos en la jerarquía de particiones.

La aplicación del método ward supone pues, en primer lugar, el cálculo de las medias de todas las variables en cada uno de los grupos definidos y, a continuación y para cada caso, encontrar la distancia cuadrática euclídea a las medias de los grupos y sumarlas (inercia intragrupos de la unión). En segundo lugar, se vuelve a aplicar el mismo criterio pero ahora considerando a los individuos de ambos grupos como miembros de uno solo (inercia intragrupos del nuevo grupo). Finalmente, y en cada etapa o partición efectuar evalúa el incremento de estas inercias intragrupos. Cuando se encuentran aquellos dos grupos el incremento es menor, serán los que formarán el nuevo grupo a la siguiente partición a una distancia δ determinada.

Este método resulta especialmente adecuado después de haber aplicado un análisis factorial donde se derivan variables factoriales construidas sobre el principio de la acumulación de la mayor parte de la varianza total explicada de la matriz informativa original, además de ser el método que tiende a proporcionar una distribución más equilibrada de los casos entre los distintos grupos con formas de hiperesfera.

Si aplicamos el procedimiento de clasificación Ward a los datos del ejemplo usado obtienen las distancias de formación de los grupos y el dendrograma siguientes:



2.3.2. *Métodos no jerárquicos*

En los métodos no jerárquicos o de partición el conjunto de casos se divide en un número predeterminado de grupos o particiones para a continuación de forma iterativa ir reasignando los casos a los grupos hasta que algún criterio de convergencia o de finalización concluye el proceso.

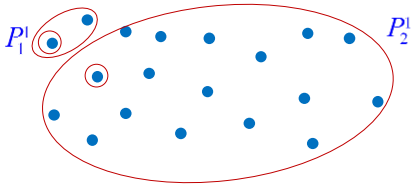
2.3.2.1. *Centros móviles*

La clasificacion no jerárquica por la agregación alrededor de centros móviles (Forgy, 1965; Diday, 1971) constituye un caso particular del procedimiento clasificatorio de nubes dinámicas y su utilización es recomendada para el procesamiento de un gran número de datos. El proceso se inicia fijando un número de grupos o particiones k, con unos centros iniciales, aleatorios o bien conocidos $(C_1^0, C_2^0, \cdots, C_K^0)$ que clasifican a las n unidades iniciales en una partición de K grupos $(P_1^0, P_2^0, \cdots, P_K^0)$. De cada unidad U_i se evalúa la distancia, cuadrática euclidiana en particular, a cada uno de los centros y se asigna al grupo cuyo centro es más próximo.

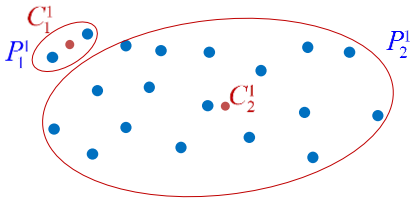
En un segundo momento se vuelven a calcular los nuevos centros de cada grupo $(C_1^1, C_2^1, \cdots, C_K^1)$ dando lugar a la partición $\{P_1^1, P_2^1, \cdots, P_K^1\}$, y así sucesivamente. El algoritmo se detendrá por el número de iteraciones que se haya fijado, por criterios de varianza explicada, bien porque iteraciones sucesivas dan lugar a la misma partición.

1. Si fijamos inicialmente 2 particiones se seleccionan al azar dos centros iniciales en la etapa 0: C_1^0 y C_2^0 por ejemplo.

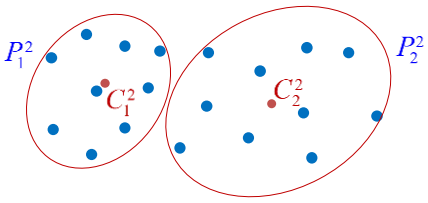
2. Se asignan los individuos al centro más cercano entre estos dos centros iniciales configurando las 2 primeras particiones P_1^1 y P_2^1.



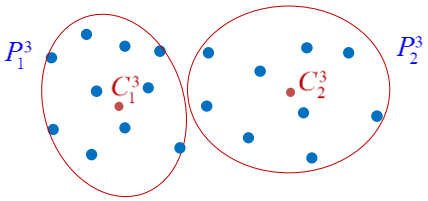
3. Con los dos nuevos grupos se calculan los nuevos centros de cada uno, los nuevos valores de esta primera etapa serán C_1^1 y C_2^1.



4. Se vuelve a asignar a cada unidad al centro más cercano como se hizo en la etapa 2, generando las particiones P_1^2 y P_2^2, en las que se calculan los nuevos centros C_1^2 y C_2^2



5. El proceso de reitera hasta la estabilización del centro de cada grupo se estabiliza.



2.3.2.2. *Método de grupos estables*

El método de grupos estables, caso particular del procedimiento de nubes dinámicas (Forgy, 1965; Diday, 1971), es especialmente adecuado cuando se dispone de un gran número de unidades a clasificar. El objetivo es construir una partición única de objetos en un número de clases determinado que se fija previamente. El procedimiento se inicia eligiendo al azar los individuos o unidades que serán los centros provisionales de un número K de clases, a continuación se asignan todos los individuos en el centro provisional más próximo. Se construye así una partición en K clases del conjunto de unidades. Esta partición en K clases se hace varias veces sobre el mismo conjunto de unidades iniciales, se consideran entonces se particiones diferentes de K iniciales.

Partición-producto

15	19	5	39
4	6	17	27
21	13	0	34
40	38	22	100

1ª Partición

2ª Partición

Si se realizan s particiones $\{P_1, P_2, ..., P_s\}$ en K clases cada una, la partición-producto, la "tabla de contingencia" de s dimensiones que resulta de cruzar las s particiones, contiene t clases, con $t=K\times s$, dando una partición $(C_1, C_2, ..., C_t)$. Las clases no vacías de la partición-producto constituyen los grupos estables.

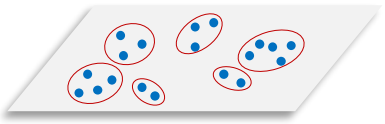
Este procedimiento permite reducir el volumen de operaciones de reagrupamiento que exige la clasificación de una gran cantidad de unidades y se obtiene un agrupamiento previo en varias clases sobre las que se puede aplicar entonces otro procedimiento clasificatorio, por ejemplo, una clasificación jerárquica.

2.3.3. *Métodos mixtos*

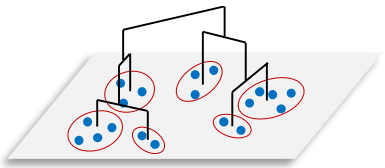
2.3.3.1. *Clasificación híbrida*

Lebart, Morineau y Piron (1995) presentan un procedimiento de clasificación con un algoritmo mixto identificado como de clasificación o aglomeración híbrida, *hybrid clustering* (Wong, 1982)[2]. Se procede en tres fases:

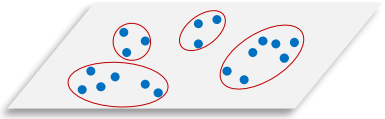
1) Se procede a efectuar una partición inicial por centros móviles para obtener unas pocas decenas o centenares de grupos homogéneos. Este procedimiento se puede combinar con el método de grupos estables cruzando diversas clasificaciones iniciales.



2) Se aplica sobre los grupos iniciales creados una clasificación jerárquica ascendente, por ejemplo, con el método de Ward, y se determina el número de clases finales a retener.



3) Finalmente la partición seleccionada se optimiza o se consolida mediante una reasignación a los diferentes grupos creados en cada partición con un nuevo proceso de clasificación por centros móviles que mejora la inercia entre los grupos.



[2] Se corresponde con el método **SEMIS** implementado en el software SPAD.

2.3.3.2. *Clasificación en dos fases*

El procedimiento de análisis de clasificación en dos fases, o clústeres bietápicos (*two step cluster analysis*)[3] es un procedimiento de exploración que busca agrupaciones naturales de un conjunto de datos desarrollado por Chiu et al. (2001).

Como su nombre indica el algoritmo de este método clasificatorio se basa dos etapas. En la primera se lleva a cabo un procedimiento similar al algoritmo de k-medias y basándose en estos resultados se aplica en una segunda etapa un procedimiento modificado de aglomeración jerárquica ascendente que combina las unidades secuencialmente para formar grupos homogéneos. Para ello se construye el llamado árbol de características de clases (*cluster feature tree*, CF).

Permite el tratamiento conjunto de variables cualitativas y cuantitativas, suponiendo que las variables son independientes, que las cualitativas siguen una distribución multinomial mientras y que las cuantitativas siguen una distribución normal[4]. El método permite especificar de entrada un número de grupos fijo, y también el número máximo, o bien la determinación automática del número de clases óptimo utilizando un criterio estadístico de selección del modelo entre diferentes soluciones posibles. Para la detección automática se pueden especificar como criterios tanto el BIC (criterio de información bayesiano) como el AIC (criterio de información de Akaike).

Mediante la construcción de un árbol de características de clases que resume los registros, el algoritmo en dos fases puede analizar archivos de datos de gran tamaño. Los casos deben estar ordenados aleatoriamente ya que el resultado clasificatorio puede depender de un orden existente.

Para determinar la similitud entre los clases que se forma se forman se emplea la medida de la verosimilitud como forma de distribución de probabilidad entre las variables. También se puede emplear la distancia euclídea si todas las variables son continuas.

2.4. Etapa 4: Clasificación y número de grupos

El número de grupos se fija con anterioridad o bien se determina a posteriori, depende de si la investigación es más hipotética o más exploratoria y de las propias características del método clasificatorio que puede exigir o no este dato inicialmente. En cualquier caso se trata de decidir y validar cuántos grupos, clases o tipos formarán parte de la clasificación o tipología final.

Se suele emplear diversos criterios para determinar el número de grupos. Milligan y Cooper (1985) constataron más de 30 técnicas distintas y Dimitriadou et al. (2002) recogen 15 índices distintos para datos binarios. Se puede consultar a Everitt (2011) para una presentación de diversos de estos criterios. Aquí destacaremos los tres siguientes:

[3] El procedimiento **TWOSTEP CLUSTER** de SPSS opera de esta forma.
[4] No obstante se considera un procedimiento robusto ante la violación de alguno de estos supuestos.

a) Distancias entre los grupos evaluadas a través de los distintos índices resultantes en cada partición en el caso de los métodos jerárquicos. La información sobre el proceso de aglomeración proporciona la información relativa al proceso de clasificación en cada etapa de partición de las unidades. El número total de etapas seguidas es igual al número de casos menos uno, y en cada etapa se calcula un valor numérico que indica la distancia a la que se forman las diversas particiones como podemos visualizar gráficamente a través del dendrograma. Cuando se produce un salto importante en esta escala es señal de que se han unido dos grupos con diferencias internas, por tanto, en el caso de decidir el número de grupos a tener en cuenta un criterio será el de retener la partición anterior a la partición que se forma con un salto importante en el índice, se "corta" el dendrograma a ese nivel. La decisión se puede tomar observando el salto en el gráfico del dendrograma, representando los coeficientes en un gráfico como el *scree test* del análisis factorial donde se disponen en abcisas el número de grupos y en ordenadas el valor del coeficiente para observar el cambio signficativo de pendiente (localizando el "codo") o realizando cálculos de diferencias para encontrar el valor de un cambio más importante[5].

b) Proporción de varianza explicada por cada partición. La partición en un número de grupos que lleve asociada una proporción de varianza explicada significativamente mayor en relación a otra partición con menor número de grupos, o cuando el aumento de varianza explicada es poco importante, determina el número adecuado. La técnica del *scree test* nos permite determinar el número de grupos localizando el cambio de pendiente que se observa en la curva que se describe con las particiones en abscisas y el porcentaje de varianza explicada en las ordenadas. Este análisis de varianza se puede aplicar siguiendo el criterio de razón de varianzas de Calinski y Harabasz (1974).

c) El criterio teórico. Cualquier decisión sobre el número de grupos debe estar acompañada y reforzada por un criterio teórico o conceptual, de interpretabilidad, que justifica y da sentido al contenido de los grupos que resultan en una agregación pertinente. Este aspecto enlaza directamente con el proceso necesario de validación de los resultados de un análisis de clasificación donde los aspectos sustantivos deben ser primordiales.

A pesar de que el resultado final suele ser una tipología con un número de grupos determinado, cabe matizar que la decisión del número de grupos no es incompatible con analizar, ampliar y presentar los resultados de un análisis clasificatorio en términos de más de una tipología. Esto es especialmente factible cuando se aplican métodos jerárquicos. Lo ilustraremos en el ejemplo sobre la segmentación del mercado de trabajo (López-Roldán, 1996b) donde la dualidad del mercado de trabajo permite hablar de una tipología con dos tipos: segmento primario y segmento secundario. Pero a su vez una tipología con tres categorías es tipificadora o explicativa de la realidad

[5] Más adelante veremos un ejercicio práctico donde se aplica una forma sencilla de concretar este razonamiento. Consiste en analizar el crecimiento del coeficiente o índice de agregación que proporciona el software estadístico al aplicar los procedimientos jerárquicos. Disponiendo los sucesivos coeficientes para cada etapa se calculan primero las diferencias de sus valores de una etapa en relación a la anterior, es decir, las diferencias primeras (la velocidad del cambio) y se repiten la operación para calcular la variación de la variación, es decir, las diferencias segundas (la aceleración del cambio). El mayor valor obtenido de las diferencias segundas coincidirá con la selección de la etapa que determina una partición con el número de grupos decidido. Para realizar estos cálculos acompañamos el texto de una hoja de cálculo con el nombre NGruposACL.xlsx.

laboral al diferenciar dentro del segmento primario entre segmento primario dependiente e independiente como la literatura ha puesto de manifiesto. E incluso una tipología con cuatro tipos donde junto a los segmentos dependiente e independiente se concibe otro de carácter intermedio dado el contexto particular de un mercado de trabajo local, es un resultado también posible que matiza, detalla y acompaña las otras dos tipologías en dos y tres tipos, y su presentación conjunta proporciona una descripción y explicación más completa y rica de la realidad social investigada.

2.5. Etapa 5: Validación e interpretación de los resultados

Validar los resultados de un análisis clasificatorio es imprescindible dado el carácter exploratorio y heurístico de la técnica multivariable y la diversidad de soluciones posibles que se pueden obtener si cambiamos el método o la medida de proximidad. Si combinamos el análisis de clasificación con un análisis factorial previo se pueden introducir variaciones adicionales en función de las decisiones en número y contenido de los factores.

Para validar los resultados se han propuesto varios criterios. Destacamos los siguientes sin que entremos en ellos en detalle (Fernández, 1991):

a) Coeficiente de correlación cophenético de Goodman y Kruskal. Aplicado habitualmente en bioestadística el coeficiente es una medida de cuán fielmente un dendrograma o la estructura jerárquica que expresa representa las verdaderas distancias, es un indicador de la posición de la formación o unión de los distintos grupos a través del cual podemos realizar comparaciones entre distintos dendrogramas resultantes de la aplicación de diversos métodos de clasificación, a la vez que nos sirve para fijar una medida de ajuste entre los datos de partida y la estructura del dendrograma.

b) Coeficiente de pertenencia. Diseñado para ver las diferencias entre los grupos según los ítems contenidos en ellos. Se calcula el cociente entre la media de la intercorrelación entre las unidades de un mismo grupo y la media de la intercorrelación de pares de ítems, uno de los cuales pertenecerá al grupo de interés. Cuando el cociente supera la unidad, y especialmente el valor 1,3 se tiene un grupo bien definido.

c) Replicación. Se trata de repetir el análisis de clasificación para diferentes submuestras de la población total para ver si las particiones contienen un nivel de consistencia interna. Si hay cambios significativos nos indica la existencia de algún problema en la partición, si no hay un cambio relevante entonces no nos proporciona más que la confirmación, la validación y estabilidad de la clasificación.

d) Simulación de Montecarlo. Es un procedimiento costoso y complejo. A partir de generadores de números aleatorios se genera una nueva matriz similar a la de los datos originales, se efectúa un análisis clasificatorio y se comparan los resultados.

e) Comparación de métodos distintos entre sí.

f) Interpretación teórica. En última instancia se relaciona o supera los criterios anteriores al dar un contenido sustantivo a las clases obtenidas coherente con un razonamiento teórico o un conocimiento experto de un fenómeno social.

Finalmente el proceso de validación consiste en asegurar que la decisión sobre la tipología final es coherente, pertinente y estable, reflejando lo mejor posible la estructura de los datos y la interpretación conceptual de los tipos.

Para interpretar los resultados de un análisis de clasificación y caracterizar los contenidos de los tipos obtenidos, información básica para la validación teórica, es posible emplear distintas técnicas descriptivas:

- Tablas estadísticas que relacionen las tipologías obtenidas con las variables que han actuado de criterios clasificatorios (variables originales y/o variables factoriales) junto con otras adicionales disponibles que puedan ser de interés, a modo de variables independientes. La variable tipológica es cualitativa, al relacionarse con variables también cualitativas podemos realizar un análisis de tablas de contingencia o de diferencia de proporciones; al relacionarse con variables cuantitativas realizaremos análisis de comparaciones de medias.
- Las comparaciones de proporciones y de medias se pueden expresar gráficamente a través de gráficos de barras y de medias.
- Cuando se emplean variables factoriales como criterios clasificatorios se pueden representar los individuos y los centros de los grupos en el espacio factorial, incluso simultáneamente con las variables originales que generaron los factores.
- Adicionalmente cuando los las unidades o las características de los datos lo facilitan se pueden emplear mapas del territorio de estudio.

Finalmente esta descripción y caracterización se expresará en un contenido sintético que expresa la identidad de cada tipo y se relacionará con la perspectiva de análisis adoptada. Es habitual y conveniente etiquetar cada uno de los tipos obtenidos con una palabra o expresión reducida que ayude a titular significativamente sus contenidos.

3. Ejemplos de aplicación

3.1. Estratificación censal de la Región Metropolitana de Barcelona

En este apartado reproduciremos los resultados de más relevancia del análisis de clasificación realizado para construir los estratos de la muestra estratificada de la Encuesta de Condiciones de Vida y Hábitos de la Población Metropolitana de 1990. Este ejemplo nos ilustrará las características básicas de los proceso de clasificación donde también se implica un análisis de componentes principales tal y como expusimos en el capítulo anterior.

El objetivo de la estratificación, y en general de cualquier proceso clasificatorio, es la clasificación en este caso de las secciones censales en estratos que, a efectos del muestreo, de ganancia en la precisión, serán la expresión de conjuntos de secciones lo más homogéneas posible dentro de cada estrato y el más heterogéneas entre sí, según las variables / criterio que se derivan del análisis factorial previo. Esta clasificación se

efectúa sin tener en cuenta ninguna restricción de contigüidad territorial, por lo que el resultado será un mapa de secciones de la Región Metropolitana dividido en un número de estratos de diferente caracterización socioeconómica: demográfica, poblacional, ocupacional, cultural, socioprofesional.

El punto de partida es la matriz reducida que se obtiene del análisis de componentes principales, una matriz de 3509 unidades (las secciones censales) y 5 variables: cuatro variables son las puntuaciones factoriales de cada sección en el espacio de 4 factores obtenidos de la aplicación del análisis de componentes principales (variables de nombre FSC1, FSC2, FSC3 y FSC4) y una quinta variable adicional que se tuvo en cuenta, la proporción de la población de la sección sobre el municipio (ZP23), dada la finalidad del análisis muestral. Como hemos sugerido anteriormente el análisis factorial nos permite cumplir condiciones deseables de las variables en base a las cuales se realiza el proceso de clasificación: por un lado, disponemos de variables que acumulan de forma reducida la mayor parte de la varianza, siendo los factores o componentes que más diferencian o discriminan a las secciones; de otro lado, al ser variables que forman base, están incorrelacionadas entre sí. Adicionalmente las variables factoriales poseen la característica de estar estandarizadas lo que favorece también la comparación entre las unidades

El criterio de proximidad, para medir el grado de similitud o disimilitud entre las secciones censales, fue la distancia cuadrática euclidiana, medida sobre la que se construye la matriz de distancias entre secciones censales de orden 3509×3509, imposible de reproducir en estas páginas dada su extensión y su escaso valor informativo ante nuestra incapacidad de captar la estructura de la información que expresa y que la técnicas clasificatoria nos porporcionará.

El proceso de clasificación seguido se dividió en dos etapas básicas:

a) Una primera y fundamental consistió en la agregación de las secciones censales según un procedimiento clasificatorio jerárquico ascendente, con el que determinar el número de grupos o estratos y obtener una primera clasificación. En este proceso de hecho se utilizaron, con objetivos de validación, varios métodos clasificatorios: distancias máximas, distancia media entre grupos, distancias entre centroides, distancias medianas y ward. Este último es que se consideró finalmente como método jerárquico. Dadas sus características de procedimiento basado en el tratamiento de la inercia resulta especialmente adecuado al tratar datos que se derivan de un análisis factorial.

b) Una segunda consistió en aplicar un procedimiento clasificatorio no jerárquico de centros móviles para, a partir de los resultados de la clasificación jerárquica, y en base a los centros iniciales obtenidos en una partición en K clases o estratos, optimizar la asignación de las secciones censales cada uno de los estratos.

La información sobre el procedimiento de clasificación jerárquica ascendente se puede observar en la tabla de aglomeración (o conglomeración) que se adjunta. Esta tabla nos muestra el proceso de clasificación desde la etapa 1 a la etapa 3508 (en este caso se reproducen sólo las últimas 10 etapas), indicando en cada etapa qué dos unidades, grupos o conglomerados se unen, a qué distancia se forma la nueva partición mediante

un coeficiente, la primera etapa donde se crearon cada uno de estos grupos o conglomerados, y finalmente la etapa siguiente donde aparecerá el conglomerado que se acaba de crear.

Lo más relevante de la tabla es analizar el comportamiento del coeficiente como criterio para determinar la partición final y por tanto decidir el número de grupos de la clasificación. El coeficiente mide la distancia entre los puntos extremos de cada par de grupos similares que se combinan para formar cada nuevo grupo o conglomerado, por tanto, el mayor o menor valor resultante es un indicador, respectivamente, de la menor o mayor homogeneidad de los grupos que conforman la nueva partición. Con este criterio, comparando varios métodos jerárquicos, evaluando la varianza explicada por cada partición y realizando un proceso de replicación con la extracción de tres muestras aleatorias de 359 secciones, el número de grupos finalmente considerado fue de 8. Esta decisión se acompañó evidentemente de un análisis de interpretación sociológica de los grupos resultantes que hicieron decidir finalmente esta clasificación de las secciones censales en 8 grupos. De hecho se realizó un análisis sistemático de las clasificaciones de 15 a 5 grupos.

Tabla III.12.3. Historial de conglomeración de las primeras y últimas particiones obtenidas con el método Ward con los datos del Padrón de 1986

		Conglomerado que se combina			Etapa en la que el conglomerado aparece por primera vez		
Etapa	nº de grupos	1	2	Coeficientes	1	2	Próxima etapa
1	3508	765	1.903	0,000	0	0	691
2	3509	480	1.108	0,001	0	0	100
3	3510	1.648	1.846	0,002	0	0	545
4	3511	2.472	2.486	0,004	0	0	1.323
5	3512	475	1.243	0,005	0	0	265
6	3513	636	1.177	0,007	0	0	865
7	3514	975	1.145	0,009	0	0	1.297
8	3515	1.196	1.212	0,011	0	0	852
9	3516	1.336	1.788	0,013	0	0	842
10	3517	797	884	0,015	0	0	1.361
…	…	…	…	…	…	…	…
3.499	10	24	32	5.065,376	3.486	3.493	3.501
3.500	9	71	72	5.334,224	3.482	3.496	3.505
3.501	8	24	28	5.758,266	3.499	3.488	3.503
3.502	7	36	38	6.249,001	3.489	3.492	3.506
3.503	6	24	55	6.911,569	3.501	3.487	3.504
3.504	5	23	24	7.942,497	3.495	3.503	3.507
3.505	4	1	71	9.143,315	3.498	3.500	3.506
3.506	3	1	36	10.779,483	3.505	3.502	3.507
3.507	2	1	23	12.864,494	3.506	3.504	3.508
3.508	1	1	6	17.540,000	3.507	3.494	0

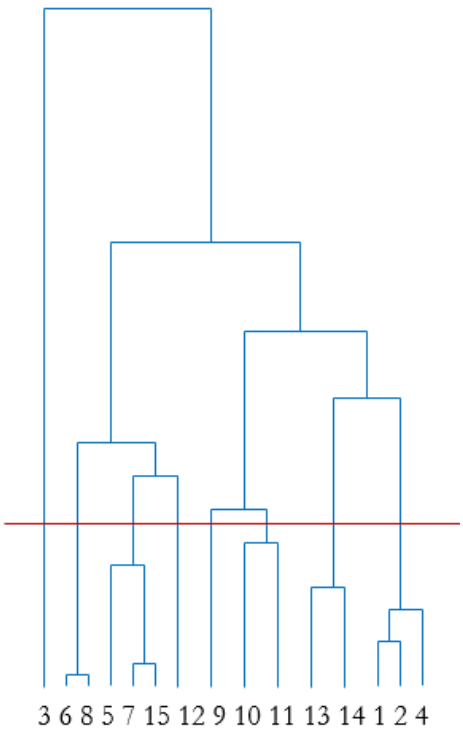
La distribución de frecuencias y el proceso de agregación con el procedimiento Ward se pueden observar en la siguiente tabla, donde se han considerado las particiones que van de 5 a 15 grupos. En negrita se han marcado los grupos que se juntan cada etapa.

Tabla III.12.4. Distribución de frecuencias de los grupos obtenidos en las particiones 15 a 5 por el método Ward con los datos del Padrón de 1986.

Número de grupos										
15	**14**	**13**	**12**	**11**	**10**	**9**	**8**	**7**	**6**	**5**
68	68	**68**	**116**	251	251	251	251	251	251	251
48	48	**48**	51	51	51	51	51	51	51	51
51	51	51	**135**	546	546	**546**	1119	**1119**	**1303**	**1588**
135	135	135	546	285	285	285	285	285	**285**	1276
546	546	546	285	573	573	**573**	**605**	1276	1276	343
145	285	285	573	605	605	605	**671**	**184**	343	
310	**310**	573	605	335	**335**	671	184	343		
140	605	605	335	336	**336**	184	343			
605	335	335	336	184	184	343				
335	336	336	184	**91**	343					
336	184	184	91	**252**						
184	91	91	252							
91	252	252								
252	**263**									
263										

No podemos disponer del árbol de conglomeración o dendrograma completo, irrepresentable dadas sus extensiones. La distribución de frecuencias que acabamos de ver nos ilustra en parte esta representación para las últimas particiones. No obstante, si consideramos estas últimas etapas del proceso de clasificación podemos representar el dendrograma de las últimas particiones. En el gráfico adjunto se han considerado 15 grupos que se agregan hasta formar uno solo.

Gráfico III.12.2. Dendrograma de las últimas particiones obtenidas por el método Ward con los datos del Padrón de 1986



Con la clasificación con el procedimiento Ward en 8 tipos o estratos se trata de aplicar un segundo procedimiento clasificatorio destinado a optimizar la asignación de las secciones a los estratos. Aplicamos el método de centros móviles a partir de los centros iniciales definidos por los 8 estratos obtenidos con Ward:

	Secciones		Centros iniciales				
Estrato	n	%	Factor 1	Factor 2	Factor 3	Factor 4	ZP23
1	251	7,2	-0,18	0,32	1,54	0,47	1,25
2	51	1,5	0,31	-0,01	6,19	0,07	7,20
3	301	8,6	1,70	0,52	-0,20	-0,92	-0,27
4	1.194	34,0	0,28	-0,44	-0,23	-0,35	-0,26
5	528	15,0	-0,87	0,98	-0,16	-0,70	-0,10
6	523	14,9	-1,02	0,33	-0,21	0,49	-0,18
7	184	5,2	-0,52	-2,29	-0,08	-0,08	-0,27
8	477	13,6	0,58	0,06	-0,34	1,47	-0,21
Total	3.509	100,0	0,00	0,00	0,00	0,00	0,00

Una vez aplicado el procedimiento clasificatorio de optimización se obtiene la distribución final de secciones por estrato y los centros finales:

	Secciones		Centros finales				
Estrato	n	%	Factor 1	Factor 2	Factor 3	Factor 4	ZP23
1	141	4,0	-0,02	0,22	2,13	0,61	1,69
2	53	1,5	0,31	0,00	6,12	0,06	7,10
3	400	11,4	1,66	0,71	-0,29	-0,42	-0,27
4	806	23,0	0,51	-0,53	-0,17	-0,50	-0,25
5	651	18,6	-0,89	0,97	-0,11	-0,64	-0,04
6	657	18,7	-0,87	0,12	-0,22	0,33	-0,17
7	326	9,3	-0,43	-1,96	-0,07	-0,24	-0,25
8	475	13,5	0,43	0,08	-0,28	1,59	-0,18
Total	3.509	100,0	0,00	0,00	0,00	0,00	0,00

Por último, una vez se dispone de la clasificación final, se trata de presentar las características que identifican cada grupo o tipo resultantes. En el caso que nos ocupa hablamos de estratos muestrales pero también de estratos sociales, e incluso de zonas sociales del territorio de la Región Metropolitana.

Para describir los estratos que se obtienen disponemos de la Tabla III.12.5 donde se muestran para cada estrato y para el total de la Región Metropolitana las medias de los de las variables factoriales y de las variables originales del Padrón de Habitantes. Se han destacado en la tabla las mayores y menores medias de cada variable. Los comentarios siguientes se destinan a destacar los rasgos identificativos más relevantes.

El estrato 4, el grosor de la región, la media urbana, se trata de un estrato con una población caracterizada por tener valores que se acercan en la mayor parte a una hipotética población de valores medios.

El estrato 2 corresponde a los pequeños municipios; ocupa una posición sociológica intermedia dentro de la región para la gran parte de las variables pero muy diferenciado en relación a la variable proporción poblacional entre la sección censal y el municipio correspondiente, se trata de secciones censales en municipios pequeños y otras variables en correlación con esta característica.

El estrato 1 es la versión atenuada del estrato 2. El que lo diferencia en relación a la media zonal consiste en su mayor parecido con el estrato 2 pero una mayor proximidad a los valores medios del conjunto.

El estrato 3 lo podemos calificar de maduros de alto status profesional y educativo. Los aposentados. Se trata de un estrato perfectamente caracterizado que se aleja de la media general fundamentalmente para incluir una elevada proporción de personas de edad, para tener un menor índice de inmigración, algo más de extranjeros, por el alto nivel educativo en todas las variables tomadas, por el relativo menor paro, por la más elevada proporción de personas con una categoría profesional alta.

Tabla III.12.5. Descripción de los estratos. Media de cada estrato y total de la Región Metropolitana por los indicadores del Padrón de Habitantes de 1986

Variable	*Media de los estratos*								*Media Total*
	Estrato 1	*Estrato 2*	*Estrato 3*	*Estrato 4*	*Estrato 5*	*Estrato 6*	*Estrato 7*	*Estrato 8*	
Factor 1	-0,18	0,31	**1,70**	0,28	-0,87	**-1,02**	-0,52	0,58	**0,0**
Factor 2	0,32	-0,01	0,52	-0,44	**0,98**	0,33	**-2,29**	0,06	**0,0**
Factor 3	1,54	**6,19**	-0,20	-0,23	-0,16	-0,21	-0,08	**-0,34**	**0,0**
Factor 4	0,47	0,07	**-0,92**	-0,35	-0,70	0,49	-0,08	**1,47**	**0,0**
ZP23 *Sección/Municipio*	1,25	**7,20**	-0,27	-0,26	-0,10	-0,18	**-0,27**	-0,21	**0,0**
P1 Menores de 15 años	24,1	20,2	14,4	16,2	24,7	**25,5**	**13,2**	23,6	**20,2**
P2 Mayores de 65 años	10,9	13,2	18,7	16,6	8,1	**7,9**	**23,2**	9,8	**13,2**
P3 Indice envejecimiento	50,9	71,0	151,5	121,8	**36,4**	33,5	**191,9**	47,5	**86,1**
P4 Inmigración	33,3	**18,7**	24,2	34,3	48,5	**48,8**	36,5	33,2	**37,4**
P5 Extranjeros	1,7	1,3	3,8	2,0	**1,2**	1,3	2,7	**2,5**	**2,0**
P6 Nuevos residentes	8,3	**12,0**	4,8	4,8	**3,4**	6,8	5,0	9,3	**5,8**
P7 Analfabetos >10 años	5,4	3,3	**1,3**	4,1	7,7	**8,7**	7,8	2,8	**5,2**
P8 Titulados medio-superior	6,2	8,1	**30,4**	10,2	2,6	**2,0**	5,1	15,1	**9,7**
P9 Escolarización 14-24	40,7	41,1	**70,6**	51,1	40,4	**36,1**	38,1	53,3	**47,7**
P10 Escolarización 2-5 años	47,5	47,9	**52,2**	42,0	**31,8**	33,9	38,0	44,3	**40,7**
P11 Parados antes ocupados	11,7	**8,3**	9,3	14,8	16,7	**18,6**	21,3	12,5	**14,9**
P12 Paro busca 1er empleo	8,1	5,9	6,2	8,1	**13,4**	12,9	9,0	**5,7**	**9,2**
P13 Paro total	19,8	**14,2**	15,5	23,0	30,2	31,5	**30,3**	18,1	**24,1**
P14 Activas >15 años	31,8	27,7	32,5	30,8	**28,0**	31,9	28,4	**39,5**	**31,8**
P15 Profesiones altas	13,2	15,2	**43,4**	20,2	7,9	**7,2**	12,4	25,4	**18,1**
P16 Profesiones bajas	5,3	2,9	**1,4**	3,3	**7,4**	6,0	3,9	2,6	**4,2**
P17 Terciario medio-comercio	12,4	11,9	**9,0**	13,2	9,3	12,3	**21,8**	12,3	**12,4**
P18 Terciario alto-finanzas	2,7	2,5	**7,4**	5,5	1,5	**1,4**	3,5	6,9	**4,3**
P19 Agropecuario	4,9	**14,8**	0,2	0,3	0,8	0,7	0,8	0,3	**1,0**
P20 Veh. privado trabajo	51,4	**59,5**	42,7	32,3	38,8	40,0	**20,2**	41,3	**37,7**
P21 Veh. privado estudio	12,2	**32,7**	13,1	4,8	2,3	**2,1**	2,4	7,8	**5,9**
P22 Veh. trabajo + estudio	33,9	**48,6**	29,7	21,0	20,7	21,7	**13,7**	27,1	**23,6**

El estrato 8 son los que podemos llamar como los emergentes. Es una versión parcialmente simétrica del estrato 4 tomando la edad como criterio diferencial. Lo que caracteriza a este estrato y lo diferencia básicamente de lo anterior es que se trata de

una población mucho más joven y más joven que la media total, así como la mayor proporción de nuevos residentes llegados en los últimos cinco años.

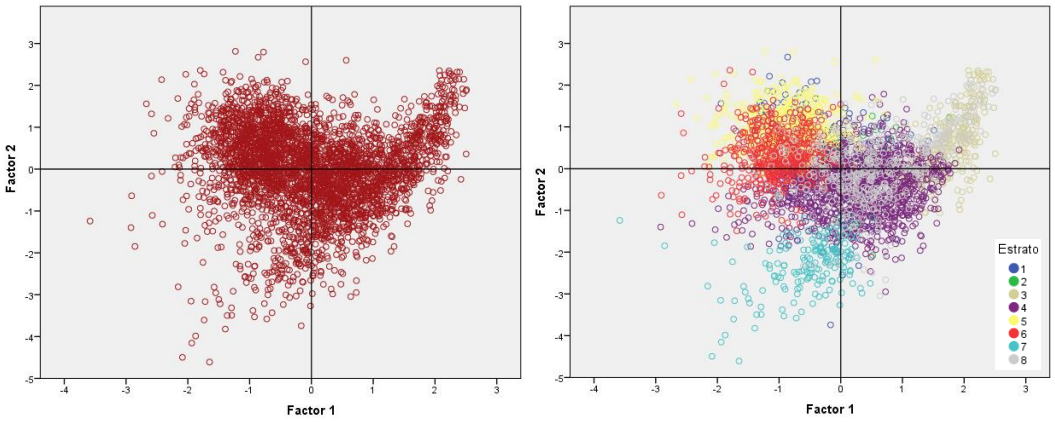
El estrato 7 son los declinantes, los mayores en los viejos comercios. En relación al estrato 4 es una versión opuesta con respecto al status social, y se caracteriza por ser el estrato donde está presente la población de mayor edad. Adicionalmente muestra una alta proporción de personas ocupadas en el sector del terciario medio, comercio y hostelería.

El estrato 6 nos sugiere la emergencia de bolsas de discriminación social. Es un estrato absolutamente opuesto al estrato 4 para la categoría socioprofesional, por el nivel educativo, ocupacional y en cierta medida por la edad.

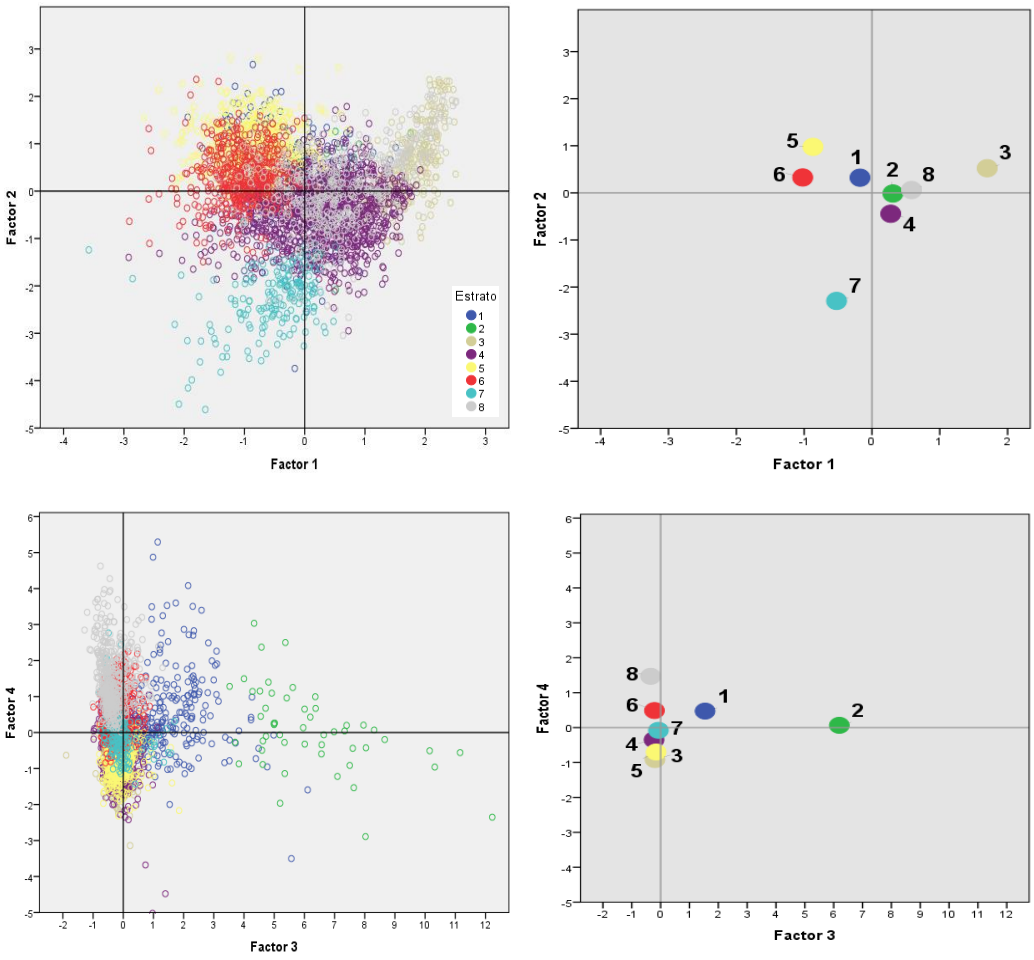
Finalmente el estrato 5 podemos decir que representa el inicio de la incertidumbre social. Es un estrato similar al sexto aunque acentuando y amortiguando algunos de sus rasgos, se trata además de una población mucho más joven y de menos parados antes ocupados, pero de mayor proporción de parados en busca del primer empleo que el estrato precedente.

Junto con la descripción de los a través de las tabla de medias se pueden realizar diversas representaciones gráficas que nos sitúan las distintas secciones según el estrato al que pertenecen. Un primer tipo de gráfico son los gráficos factoriales.

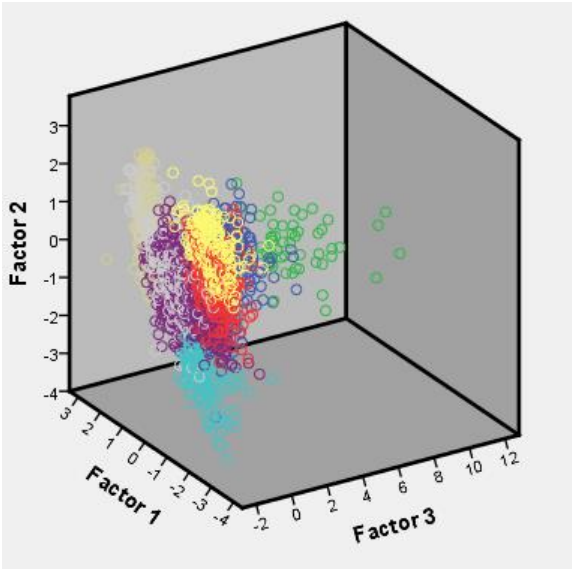
Los gráficos siguientes representan las unidades en el espacio de los dos primeros factores: el primero sin distinción del estrato y el segundo identificando con con colores el estrato de pertenencia:



El conjunto de puntos de cada estrato puede caracterizarse igualmente por los centros de éstos (el valor de la media en cada factor). En los gráficos siguientes se puede ver la correspondencia entre las nubes de puntos de las secciones, en los ejes factoriales 1 y 2 primero y 3 y 4 después, y la representación del centro de cada grupo de secciones de los estratos.



Por último se adjunta el gráfico tridimensional con los tres primeros factores.



Una segunda alternativa de representación, dada la referencia espacial de cada unidad de análisis, la sección censal, son los mapas georreferenciados. No disponemos de los mapas para los datos del Padrón del año 1986 que hemos visto. En su lugar presentamos los correspondientes a un análisis muy similar de estratificación con los datos del Censo de Población de 2001 para el conjunto de Cataluña utilizado para la construcción de la muestra de la Encuesta de Condiciones de Vida y Hábitos de la Población de Cataluña de la edición de 2006 (López-Roldán y Lozares, 2008).

En este caso se amplió el número de variables utilizadas a 82, referidas a 5222 secciones censales del conjunto del territorio catalán. El análisis de componentes principales redujo las 82 variables en términos de 7 factores principales que de forma sintética relacionamos a continuación (entre paréntesis se da el porcentaje la varianza explicada por cada factor):

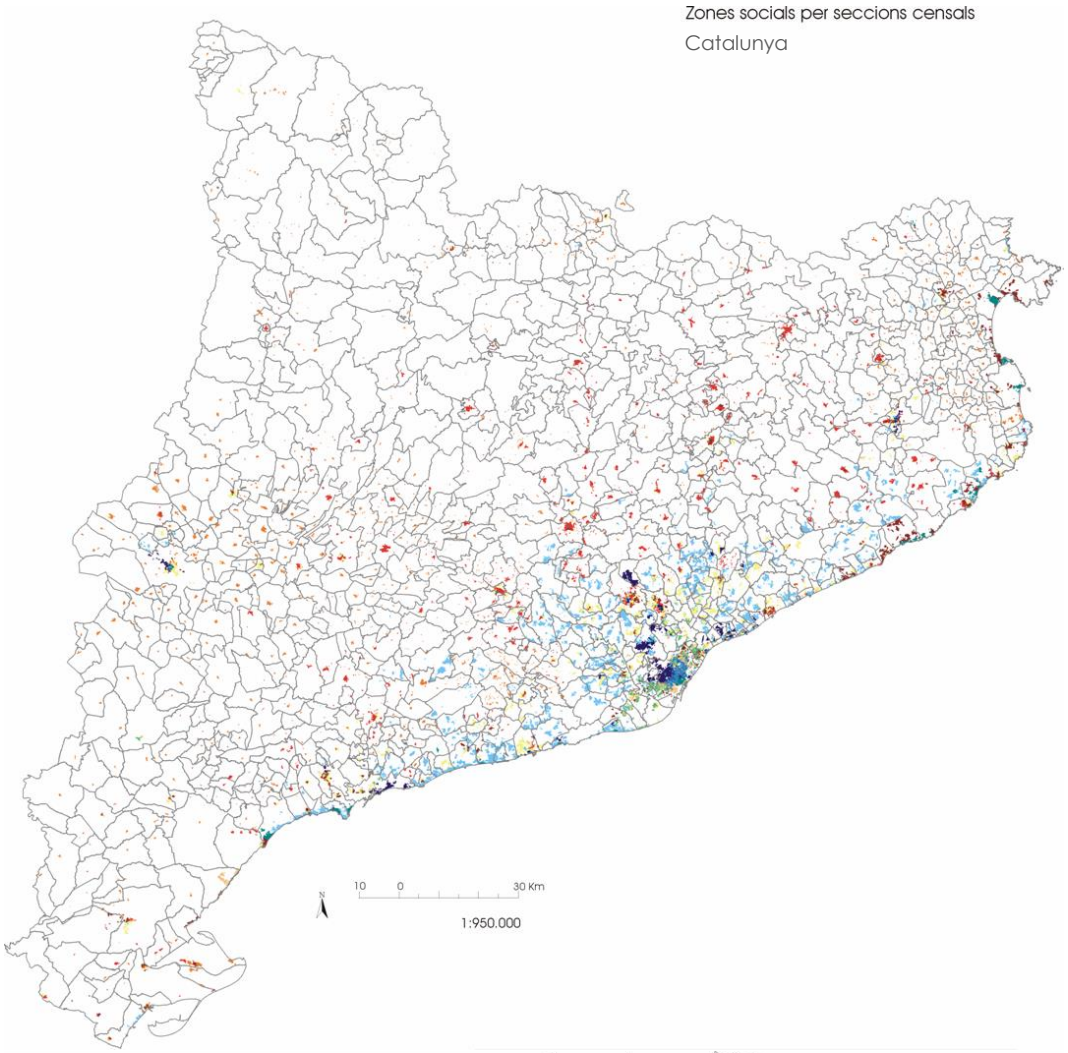
1) La categoría socio-profesional (23%)
2) Origen: autóctonos vs. inmigración (16%)
3) El ciclo vital (15%)
4) El factor rural-urbano (13%)
5) La nueva inmigración (13%)
6) Factor de actividad laboral (10%)
7) La movilidad territorial (10%)

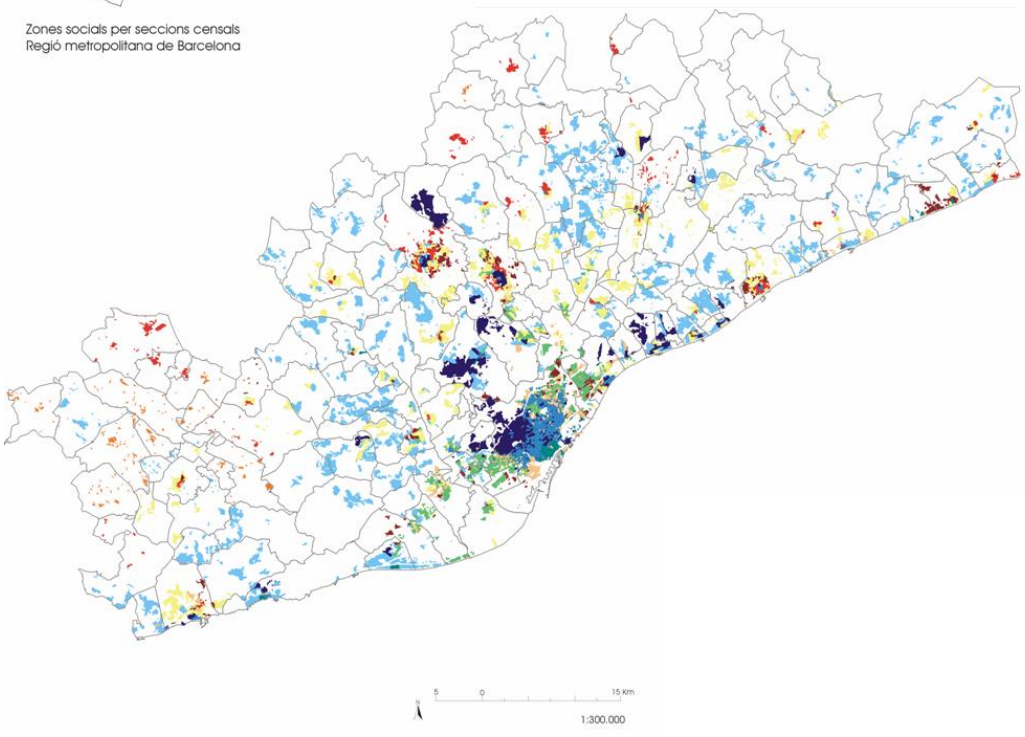
Estos factores, tras el proceso de clasificación, dieron lugar a 10 estratos con la siguiente etiqueta sintética de identidad (presentados con el color con el que aparecen en el mapa y entre paréntesis el porcentaje de secciones censales de cada estrato):

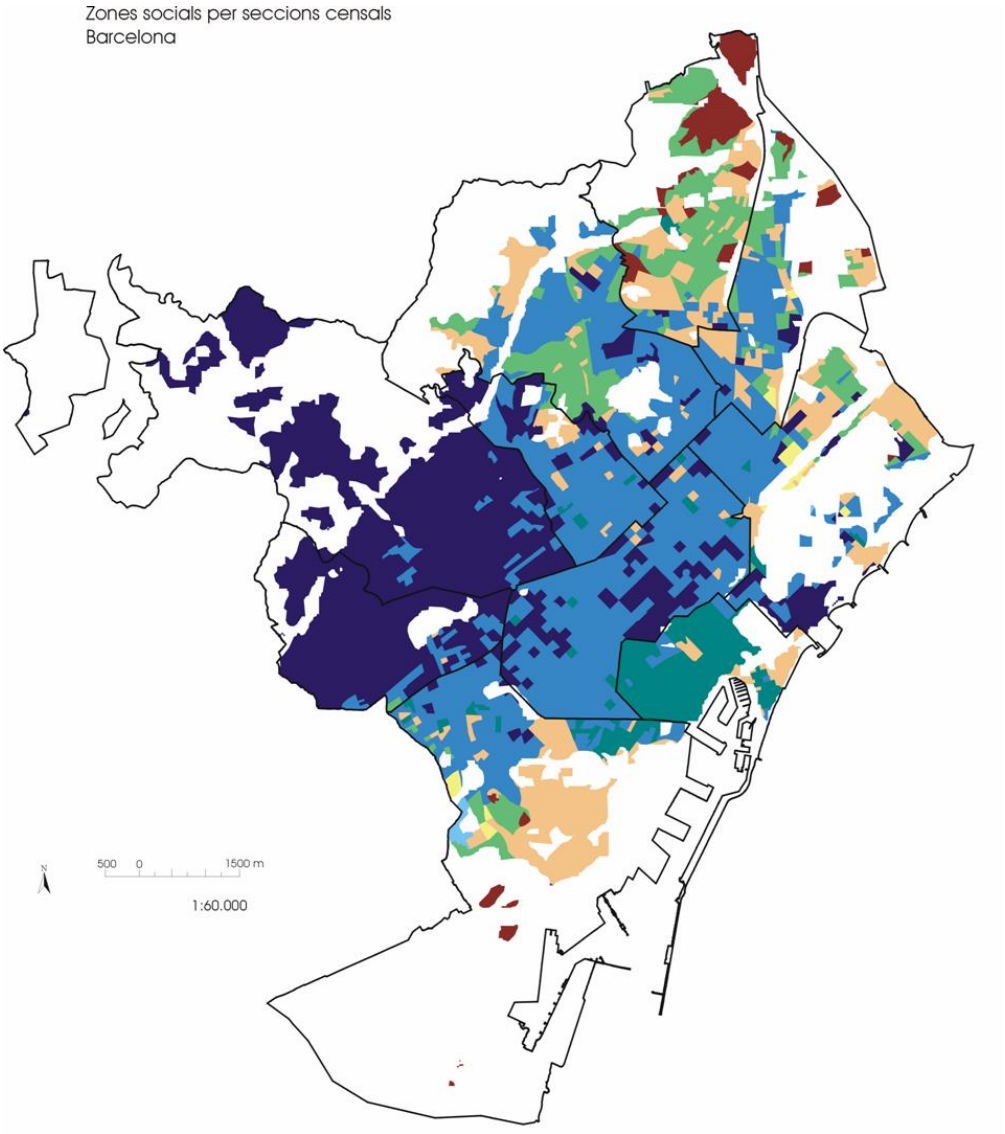
Estrato 1: Población catalana envejecida y de clases trabajadoras en ciudades no metropolitanas (11%)
Estrato 2: Movilidad de familias jóvenes de clase media (8%)
Estrato 3: Población joven de clase trabajadora de municipios urbanos (14%)
Estrato 4: Los pequeños municipios rurales (10%)
Estrato 5: Población urbana más joven de clase trabajadora precaria (8%)
Estrato 6: Clases sociales medias y altas de las grandes ciudades (8%)
Estrato 7: Antigua inmigración del área metropolitana (15%)
Estrato 8: Población envejecida urbana de la antigua inmigración (8%)
Estrato 9: Población de capital de clase media-alta con cierto envejecimiento (15%)
Estrato 10: La nueva inmigración (4%)

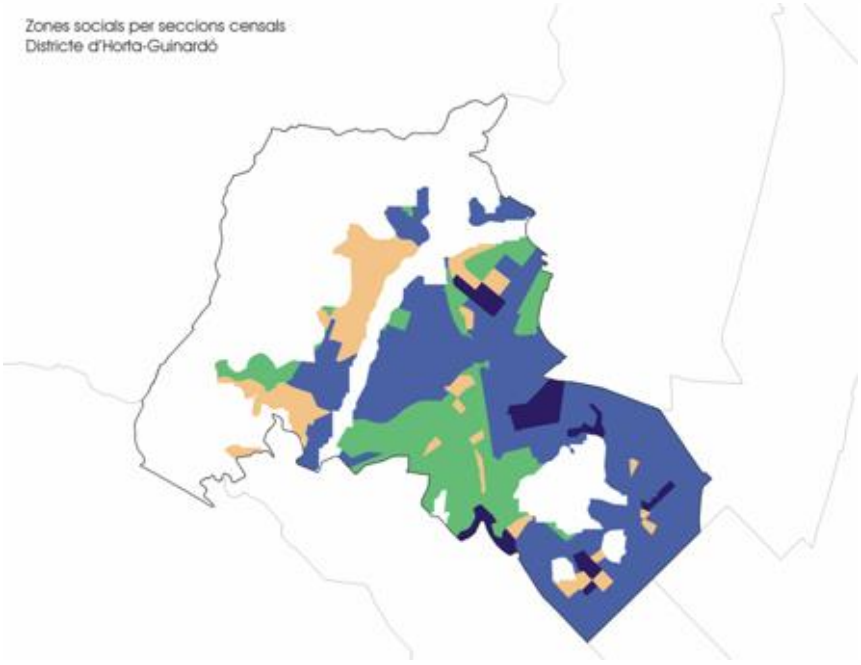
Los mapas que aparecen a continuación van desde el más general, Cataluña, a los más detallados: la región metropolitana de Barcelona, la ciudad de Barcelona y el distrito de Horta-Guinardó de la ciudad[6].

[6] Se colorea sobre el mapa el territorio de cada sección censal que corresponde a las zonas urbanizadas.


Zones socials per seccions censals
Catalunya
10 0 30 Km
1:950.000


Zones socials per seccions censals
Regió metropolitana de Barcelona
5 0 15 Km
1:300.000


Zones socials per seccions censals
Barcelona
500 0 1500 m
1:60.000


Zones socials per seccions censals
Districte d'Horta-Guinardó

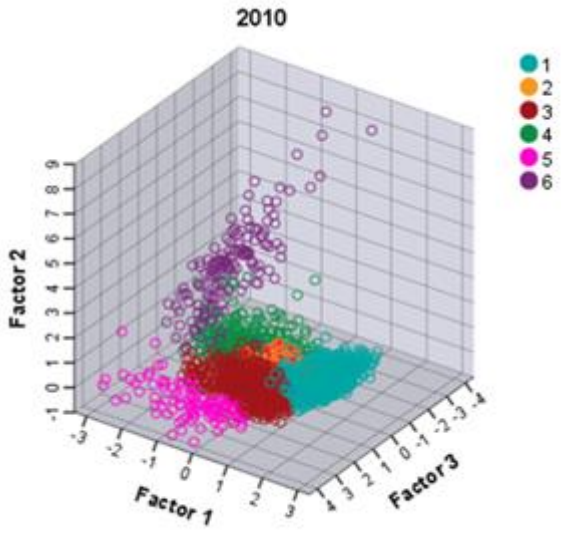
3.2. La conformación del perfil socio-espacial de la Ciudad de Buenos Aires

En el capítulo anterior presentamos la primera parte de un análisis destinado a dar cuenta de la estructura socio-habitacional de la Ciudad de Buenos Aires inspirado en el modelo del Arquitecto Torres que ha sido reducido para adaptarlo a un esquema comparativo de tres censos[7]. Una vez establecidas las principales dimensiones de diferenciación de la ciudad a partir de haber aplicado el Análisis de Componentes Principales continuamos aquí el análisis procediendo a la clasificación de los radios censales.

Tal como se ha comentado el objetivo de la clasificación consiste en la agrupación de las unidades espaciales en estratos que tipifican la realidad social del territorio a partir de conglomerados de radios lo más homogéneos posible dentro de cada grupo y con el máximo de heterogeneidad entre ellos, utilizando el método Ward. Teniendo en cuenta los saltos del árbol de agregación y el contenido de los grupos obtenidos, se concluye que la conformación óptima es la tipología con 6 grupos o tipos.

Para dar cuenta de la identidad de los grupos obtenidos se presentan a continuación diversos elementos. El gráfico siguiente representa los radios en el espacio tridimensional de los factores retenidos, donde se pueden apreciar a grandes rasgos las posiciones de todos los radios en el espacio factorial según el grupo o tipo de pertenencia de la tipología.

Gráfico de la distribución de radios en el espacio tridimensional según la tipología



Los distintos grupos o tipos se ordenan de 1 a 6 de forma decreciente, según el nivel de "estratificación socio-espacial", variable tipológica que incorpora simultáneamente

[7] Para conocer el trabajo completo que analiza la evolución de la ciudad a partir del análisis de tres censos se puede consultar "Trazando el mapa social de Buenos Aires: dos décadas de cambios en la Ciudad", *Población de Buenos Aires. 2015*, vol. 12, nº 21, p. 7-39.
http://ddd.uab.cat/pub/artpub/2015/132095/pobbueair_a2015n21p7iSPA_postprint.pdf.

las dimensiones de nivel socioeconómico, nivel de segregación y nivel de estabilidad residencial. Desde el punto de vista de la distribución de los grupos el 1 representa el 23,6%, el 2 18,3%, el 3 reúne el 28,2% de los radios y el 4 el 22,5%. Por su parte los grupos menos numerosos y más pobres, el 5 y 6, pesan 3,2 y 4,2% respectivamente. Profundizamos a continuación estos guarismos analizando la composición interna de los grupos a partir de los datos de la tabla adjunta y el mapa de estratificación de la Ciudad de Buenos Aires.

Descripción de los tipos. Media de las variables originales y de los factores según la tipología de estratificación socio-espacial de los radios censales

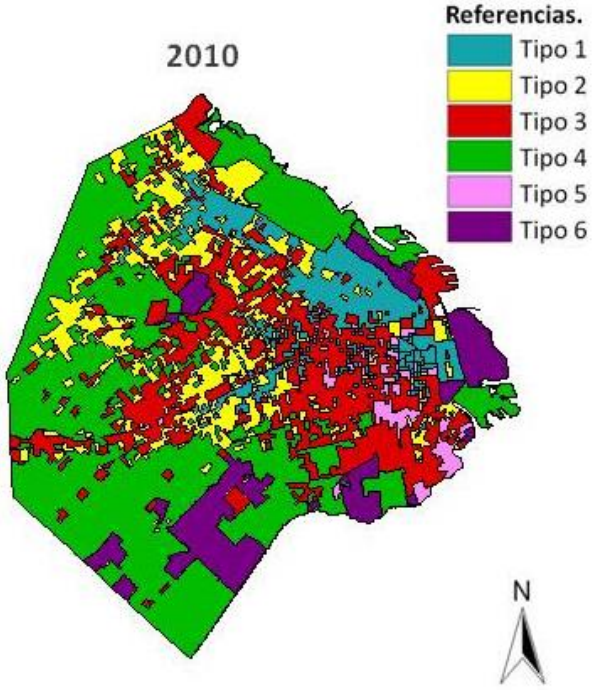
Censo 2010	**Tipo 1**	**Tipo 2**	**Tipo 3**	**Tipo 4**	**Tipo 5**	**Tipo 6**	**Total**
Factor 1	1,289	0,199	-0,137	-1,037	-0,814	-1,019	0,0
Factor 2	0,18	-0,331	-0,322	-0,229	-0,119	3,945	0,0
Factor 3	-0,162	-0,475	0,716	-0,814	3,072	0,207	0,0
Hacinamiento <0,5	42,4	35,3	30,4	27,2	19,8	9,1	32,2
Hacinamiento 1,5-2	3,4	4,5	5,9	6,9	6,0	9,9	5,4
Hacinamiento 2-3	5,0	4,7	10,1	6,8	22,6	35,1	8,6
Hacinamiento +3	0,5	0,5	1,6	0,9	5,1	11,8	1,5
Extranjeros	10,7	8,1	13,2	9,8	21,2	45,5	12,5
Nunca asistió	0,5	0,6	0,9	1,1	1,5	4,3	1,0
Primarios	13,2	17,6	21,2	26,6	27,6	43,1	21,0
Superior	11,2	11,9	11,0	9,5	8,3	1,9	10,4
Universitarios	40,3	31,2	25,5	18,7	17,5	3,2	27,3
Departamento	93,6	76,7	76,1	39,6	59,4	10,8	68,9
Inquilinato	0,5	0,5	4,8	0,7	26,0	10,0	3,0
Rancho	0,0	0,1	0,1	0,2	0,3	3,6	0,3
Inquilino	29,6	24,4	38,0	20,7	58,7	27,7	29,9
Propietario	56,5	63,2	47,5	67,9	27,6	45,7	56,4
Densidad	46.658	22.720	23.549	12.816	23.990	48.544	27.490
Nº de radios	838	651	1.003	799	113	148	3.552
% de radios	23,6	18,3	28,2	22,5	3,2	4,2	100,0

Fuente: Fachelli, Goicoechea y López-Roldán (2015)

El Tipo 1 (Tipo socio-espacial alto) se corresponde con el nivel más alto de la estratificación socio-espacial de la ciudad, con un perfil de zona residencial de clase alta. Viene caracterizado por la mayor presencia de población con estudios universitarios (40%) o superior (11%), junto con los más bajos niveles de hacinamiento, que reside en departamentos (94%) en las zonas más densamente pobladas de la ciudad.

El Tipo 2 (Tipo socio-espacial medio alto) tiene también rasgos de la clase más acomodada pero atenuada según los diversos indicadores. Aglutina a la población de altos niveles educativos (31% de universitarios y 12% de terciario), con muy bajos niveles de hacinamiento y residente en un 77% en departamentos. Su rasgo distintivo respecto del anterior es la baja densidad y la mayor propiedad. Se distribuye en la ciudad de manera inmediatamente contigua y en torno a los radios del tipo 1, como se puede observar en el Mapa siguiente.

Mapa social de la Ciudad de Buenos Aires según tipo de estrato socio-espacial



Fuente: Fachelli, Goicoechea y López-Roldán (2015)

El Tipo 3 (Tipo socio-espacial medio de centralidad) expresa el perfil medio de la ciudad tanto en los indicadores de niveles educativos como de hacinamiento, y presenta como rasgos propios la combinación de una mayor proporción tanto de departamentos (76%) como de inquilinos (38%), reduciendo la presencia del régimen de propiedad en zonas menos densamente pobladas.

El Tipo 4 (Tipo socio-espacial medio-bajo periférico) corresponde al espacio menos densamente poblado y más extenso de la ciudad, ocupando la periferia del mapa social. Son características de este grupo las casas en propiedad sin apenas hacinamiento. El perfil social se completa con la presencia en este espacio de población con niveles de estudios intermedios y primarios.

El Tipo 5 (Tipo socio-espacial bajo en riesgo de exclusión) presenta junto con el último el más bajo nivel de la estratificación socio-espacial de Buenos Aires. En él se dan altos niveles de hacinamiento de la ciudad, un alto porcentaje de inquilinatos (26%), un bajo porcentaje de propietarios (27,6%) así como una alta presencia de población extranjera (21%), siendo un territorio donde se registran bajos niveles de estudios de la población. Atendiendo a la distribución espacial de este grupo en la ciudad, se corresponde con la población residente en los hoteles y conventillos de los barrios de La Boca, San Telmo y Barracas, que recientemente viene experimentando una retracción frente al avance de la centralidad sobre la zona sudeste.

El Tipo 6 (Tipo socio-espacial en la marginalidad) finalmente es un grupo extremo que recoge como parte del mismo los radios correspondientes a las villas de emergencia. Un perfil social, por tanto, que aglutina a la población con los más bajos niveles

educativos y que refleja los mayores niveles de hacinamiento con una alta densidad de población, destacando en particular la alta proporción de ranchos (3,6%), de inquilinatos (10%) y de población extranjera (45,5%). Este grupo se aproxima socio-espacialmente al perfil del tipo anterior, uniéndose a radios de población con condiciones de vida de pobreza algo más atemperada, por tanto, en cierta medida, integrándose en la ciudad.

3.3. Un modelo de estratificación para Argentina: los estratos sociales

Tomando en cuenta los resultados presentados en el capítulo anterior sobre estratificación social en Argentina (Fachelli, 2009, 2013), se presentan los resultados obtenidos luego de aplicar el análisis de clasificación al espacio factorial obtenido con el análisis de correspondencias previo.

Recordemos que el análisis contempló 3 ejes factoriales que describían la estratificación social de los hogares de Argentina en cda una de las 4 etapas históricas comprendidas entre 1997 y 2006. El primer factor daba cuenta de la posesión o no posesión de bienes primarios, el segundo aportaba elementos a la teoría de la heterogeneidad estructural observando hogares más vinculados a sectores formales de la economía y más tradicionales enfrentado a los de servicios, con una diferenciación interna entre trabajos muy cualificados por un lado y poco calificados del otro. La tercera dimensión daba cuenta de la relación por parte de los hogares con el mercado de trabajo (insertos o no), aportando un elemento importante a la teoría de estratificación que son los hogares con individuos activos, cuestión generalmente invisibilizada por la utilización exclusiva de trabajadores activos para la elaboración de las clasificaciones sociales.

A esos tres factores se aplicó el procedimiento de clasificación mixto o híbrido, denominado `SEMIS` en el software SPAD, compuesto de tres etapas tal y como explicamos previamente en este manual: 1) Se comienza con una partición inicial por el método de nubes dinámicas resultados del cruces de diversas particiones generadas a partir de centros aleatorios con el objetivo de reducir la cantidad de unidades de análisis (se emplearon 2 particiones de 10 clases). 2) Las clases estables que se obtienen de este primer procedimiento se agregan a continuación por un método de clasificación jerárquica ascendente según el criterio de Ward o de mínima pérdida de inercia. 3) Finalmente, se aplica un método de clasificación no jerárquico de optimización que termina de ajustar los grupos y que a su vez consta de tres pasos. Paso 1: De cada individuo se evalúa la distancia a cada uno de los centros y se le asigna al que está más próximo. Paso 2: Se calcula de nuevo el centro de cada grupo y se obtiene una nueva reasignación y, en consecuencia, una nueva partición. Paso 3: De esta manera se repite el algoritmo hasta la estabilización de los grupos.

Aplicando esta metodología se obtuvieron 4 estratos en cada uno de los años analizados. Los rasgos más relevantes que caracterizan a los hogares que componen cada grupo social nos llevaron a identificar los principales estratos sociales que describen a la población Argentina y que se presentan sintéticamente a continuación, con la tabla de distribución de frecuencias para cada año.

Tabla de Estratificación Social en Argentina (% de hogares)

Período	Estabilidad	Post Crisis	Recuperación	
			Incipiente	Consolidada
Estratos Sociales	1997	2002	2003	2006
Alto	15,3	14,0	14,5	16,2
Medio Laboral Activo	46,5	43,4	42,5	45,8
Medio Laboral Inactivo	21,2	22,3	21,3	17,9
Bajo	17,0	20,2	21,7	20,1
Total	100,0	100,0	100,0	100,0
Hogaes expandidos	6.354.293	7.115.643	6.914.843	7.245.436

Fuente: Fachelli (2009)

Estrato alto, mayormente compuesto por hogares:
- con patrones o empleadores y profesionales asalariados,
- con nivel educativo superior o universitario completo,
- sin hacinamiento, con baño de uso exclusivo y propietarios,
- con decil de ingreso per cápita familiar alto (octavo al décimo).

Estrato medio laboral activo, mayormente compuesto por hogares:
- con trabajadores formales manuales
- con educación secundaria,
- sin hacinamiento (aunque hay un porcentaje pequeño de hogares que tiene hacinamiento), con baño de uso exclusivo y propietarios (con un pequeño porcentaje de hogares que son inquilinos),
- con decil de ingreso per cápita familiar medio (cuarto al octavo).

Estrato medio laboral inactivo, mayormente compuesto por hogares:
- no vinculados al mercado de trabajo (que superan el 70% y es lo que le da el nombre a esta categoría),
- con educación primaria y en menor medida secundaria,
- sin hacinamiento, con baño de uso exclusivo y propietarios,
- perteneciente a todos los deciles de ingreso per cápita familiar aunque con mayor presencia del quinto al séptimo.

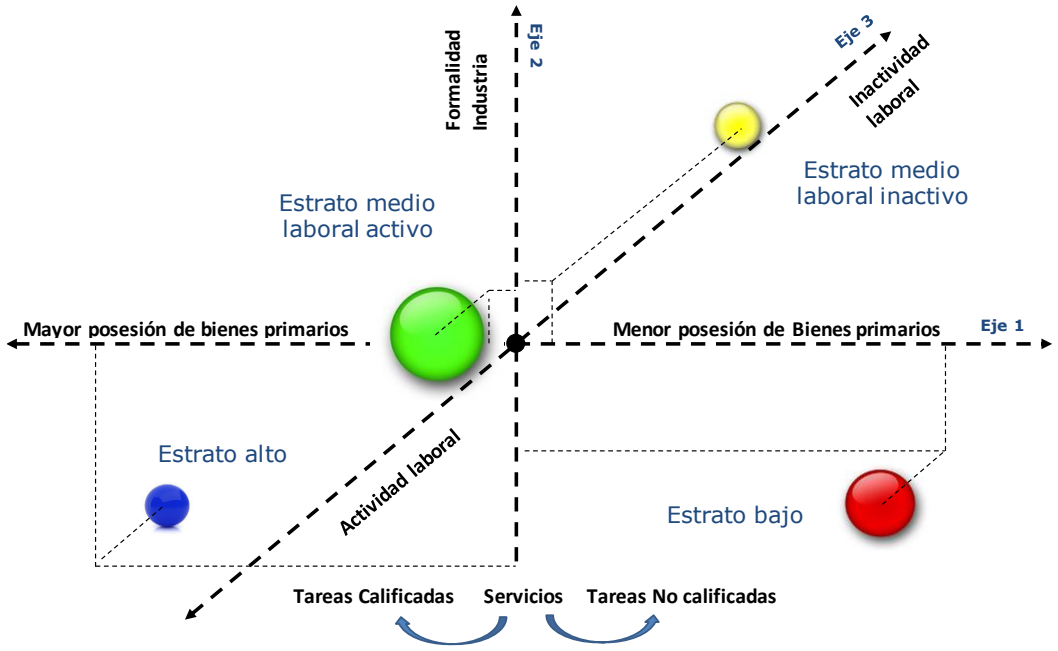
Estrato bajo, mayormente compuesto por hogares:
- con trabajadores informales, cuenta propias con calificación operativa o sin calificación y en menor medida trabajadores formales,
- con primaria completa y en menor medida secundaria incompleta,
- con hacinamiento, baño de uso exclusivo (con presencia de hogares que comparten baño o que no lo tienen) y propietarios (aunque es el estrato con mayor porcentaje de hogares que ocupan gratuitamente la vivienda),
- con bajo decil de ingreso per cápita familiar (primero al tercero).

En términos generales y desde el punto de vista de la evolución de estos estratos puede observarse que existe un "estrato medio" bastante numeroso, que en todos los casos supera el 40% de los hogares y cuyos miembros están vinculados al mercado laboral. En segundo término encontramos un estrato social conformado prácticamente por hogares que no tienen vinculación al mercado laboral y que ronda el 21%, aunque en 2006 es menor producto de la recuperación económica y la mayor cantidad de puestos

de trabajo. En tercer lugar encontramos al "estrato bajo" que crece con la crisis económica del 2002. Finalmente, tenemos el sector más pequeño de hogares, que son los pertenecientes al "estrato alto" que varía entre el 14% y 16% de los hogares argentinos.

En el gráfico siguiente se pueden observar de forma esquemática los dos elementos comentados hasta ahora, por un lado la posición que ocupan los estratos sociales según su tamaño relativo y por otro lado, su anclaje en las dimensiones de análisis o ejes factoriales (que reflejan una estructura básicamente similar en todos los años). El eje horizontal hace referencia a la dimensión 1 (acumulación/ desacumulación de bienes primarios) y el eje vertical es el correspondiente a la dimensión 2 (que discrimina entre formas tradicionales de inserción en el mercado laboral frente a otros tipos de inserción). La tercera dimensión muestra los polos opuestos entre los hogares vinculados al mercado de trabajo y el grupo de hogares relacionado con la inactividad.

Gráfico de las dimensiones de la estratificación social y posición de los estratos



Fuente: Fachelli (2013)

El cruce de los ejes representa el "hogar promedio argentino" de forma tal que los estratos más alejados de este hogar típico promedio, es, por un lado, el "estrato alto", que es un sector muy reducido comparado con el "estrato medio laboral activo" y es el que se encuentra más alejado del hogar promedio, lo que ya indica una de las características específicas que describe a ese grupo. Por el otro lado, el "estrato bajo" aunque es más numeroso que el estrato alto, también comparte la característica de estar bastante alejado del hogar promedio, pero por razones opuestas al estrato alto. Los estratos medios, tanto el laboral activo como el laboral inactivo, se ubican más cerca del cruce de los ejes.

Si comparamos la evolución por período observamos que el cambio que se produce en el tamaño de los estratos muestra que con la crisis se da una reducción de los hogares del "estrato alto" y "medio laboral activo" y un aumento significativo del "estrato bajo"

y en menor medida del "laboral inactivo" (que es el que recibe los hogares cuyos miembros quedan sin trabajo además de los hogares con inactivos). La contracara se da en el período de recuperación. Sin embargo es importante destacar que el "estrato bajo" crece con la crisis económica del 2002 y no vuelve a recuperarse.

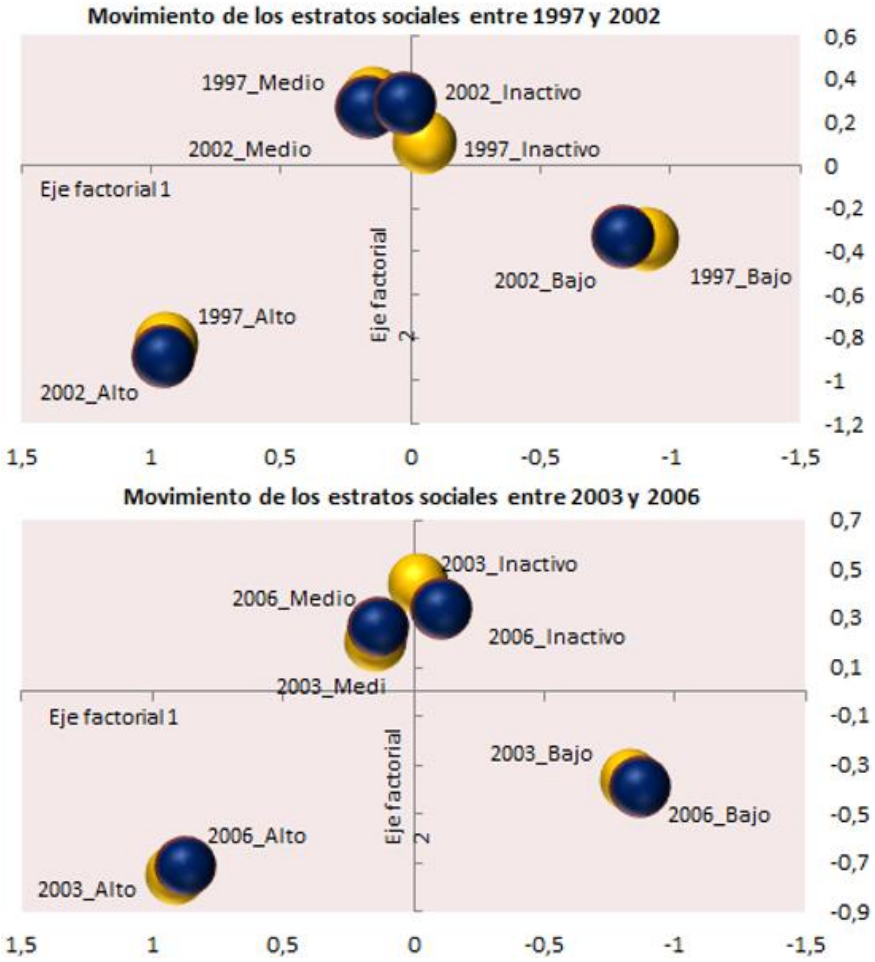
Cambio porcentual entre estratos

Estratos Sociales	Cambio % 1997 y 2002	Cambio % 2003 y 2006
Alto	-8,1	11,8
Medio Laboral Activo	-6,6	7,7
Medio Laboral Inactivo	5,1	-16,1
Bajo	19,0	-7,2

Fuente: Fachelli (2009)

El gráfico siguiente refleja el cambio mencionado, mostrando el movimiento de los estratos según la situación socioeconómica del momento analizado. Los cambios que afectan a los hogares entre 1997 y 2002 produce una trayectoria que va en un sentido, pues hay un efecto de la crisis que impacta sobre los hogares y tanto sus bienes primarios, como su relación con el mercado de trabajo se modifica de manera que se pierden activos.

Trayectorias de los estratos de hogares en el espacio factorial



Fuente: elaboración propia sobre la base de Fachelli (2009)

Por el contrario, en el período de recuperación -entre 2003 y 2006- debido a la mejora de la situación económica, la trayectoria tiene un sentido inverso a la presentada entre 1997 y 2002 captando la mejora de la situación de los hogares.

Finalmente, como conclusión del ejemplo presentado, hemos podido observar las distintas posiciones que ocupan los hogares en un espacio tridimensional producto del análisis multivariado empleado. Así el esquema no concibe los grupos de antemano, sino que organiza los hogares que más se parecen entre sí y los diferencia de aquellos que tienen otras características definitorias, y este posicionamiento implica una medición cualitativa (tipológica), que es susceptible de ser mensurada objetivamente.

4. Análisis de clasificación con SPSS

A partir del ejemplo utilizado en el capítulo anterior, donde tratamos los datos censales del municipio de Alcobendas y donde se aplicó un análisis factorial de componentes principales para estudiar de estructura social en el municipio, se trata de realizar un análisis de clasificación automática a partir de las variables factoriales obtenidas y presentar estos resultados con la ayuda del software del SPSS y así conocer su utilización.

Los resultados del análisis factorial de componentes principales nos proporcionaron dos componentes o factores con los que reducir y expresar sintéticamente las diferencias entre las secciones censales de este municipio según la información que se consideró en 15 variables censales:

	Media	Desviación típica
INMIGR % Inmigrantes	28,44	9,01
PAROJU Tasa Paro Juvenil (15-21)	12,28	4,95
PEVEN % Eventuales	14,94	5,72
PDTORE % Directores	1,31	2,18
POPER % Operarios	34,94	14,01
PTECN % Técnicos	10,75	8,95
PEVENM % Mujeres Eventuales	24,78	10,04
PTECNM % Mujeres Técnicas	13,53	8,65
PADMIM % Mujeres Administrativas	21,17	8,56
PDOMEM % Mujeres en Servicio Doméstico	25,22	10,42
PPOBJO Población -15 / +65	27,11	12,20
SINEST % Sin Estudios	25,28	9,54
SUPER % Estudios Superiores	6,03	8,77
TACTM Tasa Actividad Femenina	26,31	5,02
TPAROM Tasa Paro Femenino	27,36	8,25

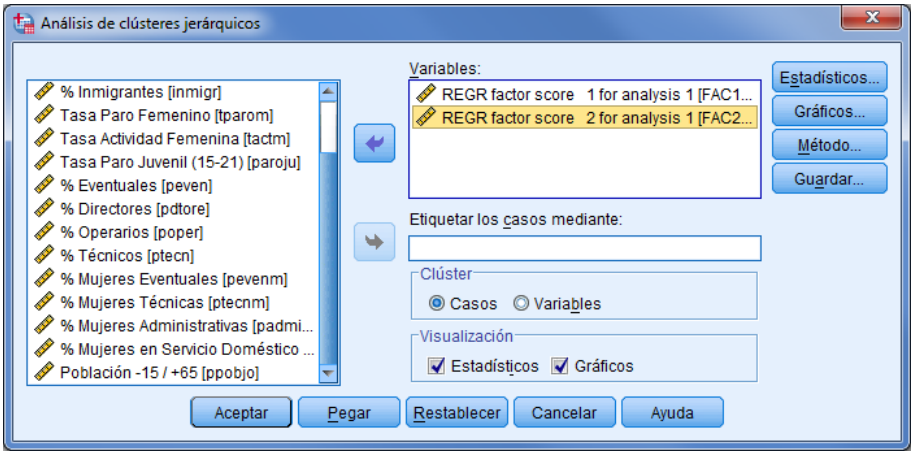
Esto supone disponer de dos variables factoriales en base a las cuales se caracterizan las secciones y a partir de las cuales podemos proceder a efectuar un análisis de clasificación automática. Estas dos variables densas actuarán de variables criterio para agrupar las secciones censales en un número reducido de grupos según sus similitudes y según un método clasificatorio determinado.

A continuación presentaremos los principales resultados de tablas y gráficos que se obtienen del análisis de clasificación, dando cuenta en primer lugar de las pautas básicas para reproducir el análisis de clasificación mediante el programas estadístico SPSS, con los cuadros de diálogo que aparecen en la ejecución de instrucciones a través del menú.

Una vez abierto el archivo de datos, en nuestro caso Alcobendas.sav, que contiene las 15 variables iniciales de caracterización de las secciones censales más las dos variables factoriales (FAC1_1 y FAC2_1), accederemos al procedimiento **CLUSTER** del SPSS que realiza análisis de clasificación de tipo jerárquico. En la barra de menús se localiza en:

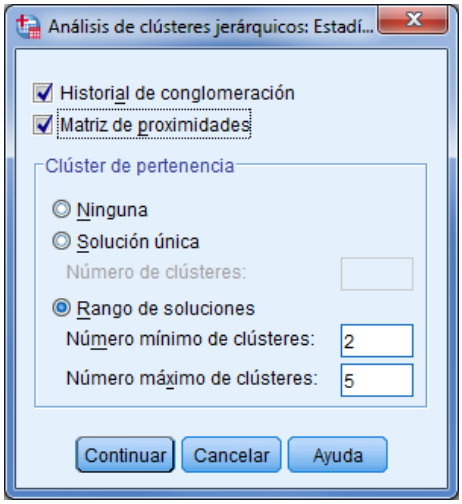
Analizar / Clasificar / Conglomerados jerárquicos

El cuadro de diálogo inicial del procedimiento de Análisis de Conglomerados Jerárquico es el siguiente:

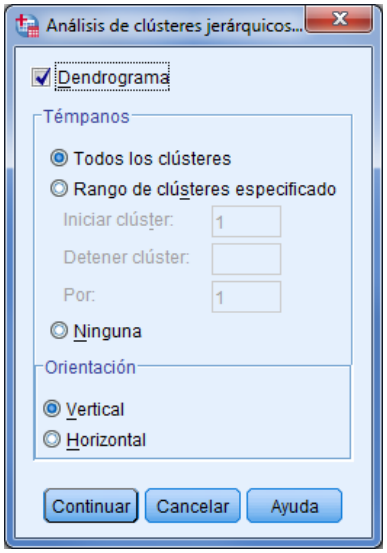


En el cuadro de Variables se han colocado las dos variables factoriales que clasificarán los 36 casos, las secciones censales. Se muestran marcadas las opciones que posibilitan mostrar estadísticos y gráficos. A continuación especificaremos las diferentes opciones de este procedimiento de análisis a través de los botones: Estadísticos, Gráficos, Método y Guardar.

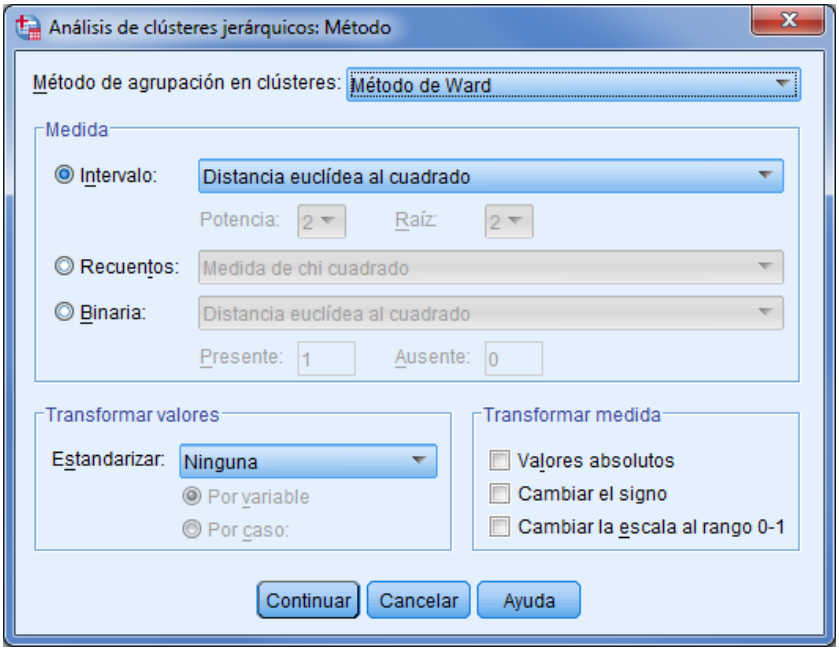
En los diferentes cuadros de diálogo que aparecerán a continuación se han marcado las especificaciones que se utilizarán de este procedimiento, y que reproducen los resultados que se presentan a continuación.



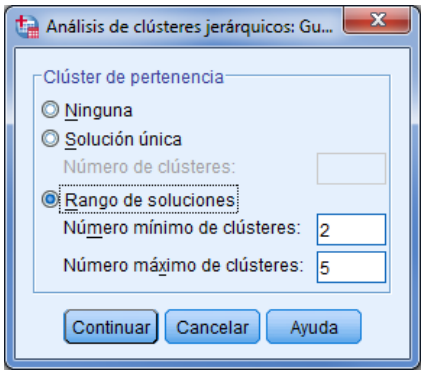
Si empezamos por las opciones de Estadísticos, marcaremos en primer lugar Historial de conglomeración para obtener una tabla con el proceso de agrupación de los casos en clases. Además de informar de cómo se combinan en cada etapa se proporciona el coeficiente que establece las distancias entre los casos o los grupos que se combinan. Igualmente pediremos la matriz de distancias pues el número de casos es reducido, en cualquier otro caso con un mayor número de individuos no resulta de interés sobre todo si se trata de unidades no identificables. En tercer lugar pediremos el listado de todos los casos con el conglomerado o grupo al que pertenecen en una clasificación en este caso con 5 grupos, con 4 grupos, con 3 grupos y con 2 grupos, es decir, en el rango de soluciones de 2 a 5.



Las especificaciones de los gráficos nos permiten obtener por un lado el Dendrograma (representación en forma de árbol jerárquico que muestra qué grupos se unen a cada etapa y a qué distancia) y el Diagrama de Témpanos, en este caso en dirección vertical y para todas las etapas, que complementa la visión progresiva de la formación de los grupos que da el dendrograma.



En relación al método de clasificación (Método de agrupación en clústeres) disponemos de un listado de siete posibles métodos de clasificación jerárquica ascendente. En el ejemplo que reproducimos utilizaremos el método Ward y considerando el uso de la Distancia cuadrática euclidiana como medida de proximidad para variables cuantitativas que exige este procedimiento clasificatorio. Dado que nuestras variables provienen de un análisis factorial no necesitamos plantearnos una transformación de valores o de medidas.



Finalmente tenemos la posibilidad de guardar como variables las diferentes clasificaciones con diferente número de grupos. De hecho podemos elegir una clasificación única o un rango de valores como hemos especificado nosotros. En este caso tendremos las 36 secciones censales clasificadas en 5, 4, 3 y 2 grupos, clasificaciones y variables que pueden ser utilizadas con otros comandos del SPSS para complementar el análisis descriptivo del proceso de agrupación y de las propias clasificaciones y decidir finalmente en la interpretación la clasificación final por la que optamos.

El procedimiento de análisis de clasificación se corresponde con el mando **CLUSTER** del lenguaje del SPSS. La mayor parte de las posibilidades que permite este comando se pueden procesar mediante las instrucciones del menú, pero la utilización del lenguaje de comandos nos facilita además: pedir los resultados de varios métodos clasificatorios a la vez en una sola ejecución del comando, leer y guardar matrices de proximidades en un fichero, o especificar los nombres de las variables de clasificación que se guardarán[8].

Los resultados de tablas y gráficos del procedimiento que se obtienen son los siguientes.

En primer lugar aparece la matriz de distancias entre las 36 secciones censales del municipio de Alcobendas.

Matriz de proximidades

Caso	Distancia euclídea al cuadrado																																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	0,0	0,0	0,2	0,7	0,3	0,2	0,1	0,1	0,0	0,5	1,7	0,1	0,1	0,1	0,0	0,2	0,2	0,2	0,1	0,9	0,7	0,0	3,4	2,4	1,2	3,9	4,1	0,2	1,0	9,7	6,7	0,4	10,9	9,0	11,9	13,4
2	0,0	0,0	0,2	0,6	0,3	0,2	0,0	0,1	0,1	0,6	1,7	0,1	0,1	0,1	0,1	0,3	0,1	0,2	0,1	0,9	0,6	0,0	3,2	2,3	1,1	3,7	4,0	0,2	0,9	9,4	6,5	0,3	10,5	8,7	11,6	13,1
3	0,2	0,2	0,0	1,5	0,8	0,0	0,3	0,0	0,0	0,2	0,8	0,1	0,4	0,5	0,4	0,0	0,1	0,1	0,1	0,4	1,2	0,2	4,4	3,5	1,9	5,0	5,6	0,7	1,2	11,4	7,9	0,9	10,2	8,2	10,1	10,8
4	0,7	0,6	1,5	0,0	0,2	1,5	0,4	1,3	1,1	2,3	4,5	1,1	0,6	0,3	0,5	1,7	1,1	1,5	1,1	3,0	0,2	0,7	1,4	0,7	0,4	1,8	1,6	0,2	0,9	5,8	4,0	0,1	11,0	9,5	14,3	17,8
5	0,3	0,3	0,8	0,2	0,0	0,9	0,1	0,6	0,5	1,6	3,2	0,7	0,4	0,1	0,3	1,0	0,4	1,0	0,7	2,2	0,1	0,4	1,6	1,0	0,3	2,0	2,2	0,0	0,4	6,5	4,2	0,0	9,3	7,8	11,7	14,7
6	0,2	0,2	0,0	1,5	0,9	0,0	0,4	0,1	0,0	0,1	0,9	0,0	0,3	0,4	0,3	0,0	0,2	0,0	0,1	0,3	1,4	0,1	4,8	3,7	2,1	5,4	5,8	0,7	1,6	12,0	8,5	1,0	11,4	9,4	11,3	11,8
7	0,1	0,0	0,3	0,4	0,1	0,4	0,0	0,2	0,2	0,9	2,2	0,3	0,2	0,0	0,1	0,5	0,2	0,4	0,3	1,3	0,3	0,1	2,5	1,7	0,7	3,0	3,3	0,1	0,6	8,2	5,5	0,2	9,8	8,1	11,4	13,5
8	0,1	0,1	0,0	1,3	0,6	0,1	0,2	0,0	0,0	0,3	1,0	0,1	0,4	0,4	0,3	0,0	0,1	0,1	0,1	0,5	1,0	0,1	4,0	3,1	1,6	4,5	5,2	0,5	1,0	10,7	7,4	0,8	9,8	7,9	10,0	11,0
9	0,0	0,1	0,0	1,1	0,5	0,0	0,2	0,0	0,0	0,3	1,2	0,0	0,2	0,2	0,2	0,1	0,1	0,1	0,1	0,6	0,9	0,1	3,9	2,9	1,6	4,4	4,9	0,4	1,1	10,6	7,3	0,6	10,5	8,6	10,9	12,0
10	0,5	0,6	0,2	2,3	1,6	0,1	0,9	0,3	0,3	0,0	0,6	0,2	0,6	1,0	0,7	0,2	0,6	0,1	0,2	0,1	2,3	0,5	6,4	5,1	3,3	7,1	7,6	1,4	2,5	14,5	10,7	1,7	13,1	10,8	12,1	11,8
11	1,7	1,7	0,8	4,5	3,2	0,9	2,2	1,0	1,2	0,6	0,0	1,2	2,2	2,5	2,3	0,7	1,3	0,9	1,3	0,3	3,9	1,7	8,6	7,5	5,0	9,4	10,7	3,0	3,5	17,6	12,9	3,5	11,7	9,4	9,0	7,6
12	0,1	0,1	0,1	1,1	0,7	0,0	0,3	0,1	0,0	0,2	1,2	0,0	0,1	0,3	0,2	0,1	0,3	0,0	0,0	0,5	1,2	0,1	4,4	3,3	1,9	5,0	5,3	0,6	1,5	11,4	8,1	0,8	11,9	9,8	12,2	13,0
13	0,1	0,1	0,4	0,6	0,4	0,3	0,2	0,4	0,2	0,6	2,2	0,1	0,0	0,1	0,0	0,5	0,4	0,3	0,1	1,1	0,8	0,0	3,5	2,4	1,4	4,0	4,0	0,3	1,4	9,8	7,1	0,4	12,4	10,4	13,7	15,2
14	0,1	0,1	0,5	0,3	0,1	0,4	0,0	0,4	0,2	1,0	2,5	0,3	0,1	0,0	0,0	0,6	0,3	0,5	0,2	1,4	0,4	0,1	2,6	1,7	0,8	3,1	3,2	0,1	0,9	8,3	5,7	0,1	10,9	9,1	12,6	14,8
15	0,0	0,1	0,4	0,5	0,3	0,3	0,1	0,3	0,2	0,7	2,3	0,2	0,0	0,0	0,0	0,5	0,4	0,3	0,1	1,2	0,7	0,0	3,2	2,1	1,1	3,7	3,7	0,2	1,2	9,2	6,6	0,3	11,8	9,9	13,2	15,0
16	0,2	0,3	0,0	1,7	1,0	0,0	0,5	0,0	0,1	0,2	0,7	0,1	0,5	0,6	0,5	0,0	0,2	0,1	0,2	0,3	1,5	0,3	4,9	3,9	2,2	5,5	6,2	0,9	1,5	12,1	8,5	1,1	10,6	8,5	10,1	10,6
17	0,2	0,1	0,1	1,1	0,4	0,2	0,2	0,1	0,1	0,6	1,3	0,3	0,4	0,3	0,4	0,2	0,0	0,3	0,3	0,9	0,7	0,2	3,2	2,5	1,2	3,7	4,5	0,4	0,6	9,4	6,2	0,6	8,6	6,9	9,2	10,7
18	0,2	0,2	0,1	1,5	1,0	0,0	0,4	0,1	0,1	0,1	0,9	0,0	0,3	0,5	0,3	0,1	0,3	0,0	0,0	0,3	1,5	0,2	5,0	3,8	2,3	5,6	6,0	0,8	1,8	12,3	8,9	1,0	12,0	9,9	11,8	12,3
19	0,1	0,1	0,1	1,1	0,7	0,1	0,3	0,1	0,1	0,2	1,3	0,0	0,1	0,2	0,1	0,2	0,3	0,0	0,0	0,5	1,2	0,1	4,3	3,2	1,9	4,9	5,2	0,5	1,5	11,3	8,1	0,7	12,1	10,0	12,5	13,4
20	0,9	0,9	0,4	3,0	2,2	0,3	1,3	0,5	0,6	0,1	0,3	0,5	1,1	1,4	1,2	0,3	0,9	0,3	0,5	0,0	2,9	0,8	7,3	6,0	4,0	8,1	8,7	1,9	3,0	15,9	11,7	2,3	13,2	10,8	11,5	10,7
21	0,7	0,6	1,2	0,2	0,1	1,4	0,3	1,0	0,9	2,3	3,9	1,2	0,8	0,4	0,7	1,5	0,7	1,5	1,2	2,9	0,0	0,7	1,0	0,6	0,1	1,3	1,7	0,2	0,2	5,2	3,1	0,1	8,0	6,7	11,0	14,6
22	0,0	0,0	0,2	0,7	0,4	0,1	0,1	0,1	0,1	0,5	1,7	0,1	0,0	0,1	0,0	0,3	0,2	0,2	0,1	0,8	0,7	0,0	3,5	2,5	1,3	4,0	4,2	0,3	1,1	9,8	6,9	0,4	11,2	9,2	12,1	13,5
23	3,4	3,2	4,4	1,4	1,6	4,8	2,5	4,0	3,9	6,4	8,6	4,4	3,5	2,6	3,2	4,9	3,2	5,0	4,3	7,3	1,0	3,5	0,0	0,2	0,5	0,0	0,4	1,9	1,2	1,6	0,7	1,6	7,1	6,4	12,7	18,8
24	2,4	2,3	3,5	0,7	1,0	3,7	1,7	3,1	2,9	5,1	7,5	3,3	2,4	1,7	2,1	3,9	2,5	3,8	3,2	6,0	0,6	2,5	0,2	0,0	0,3	0,3	0,3	1,1	1,0	2,5	1,4	0,9	8,6	7,7	13,8	19,3
25	1,2	1,1	1,9	0,4	0,3	2,1	0,7	1,6	1,6	3,3	5,0	1,9	1,4	0,8	1,1	2,2	1,2	2,3	1,9	4,0	0,1	1,3	0,5	0,3	0,0	0,7	1,1	0,4	0,3	4,0	2,2	0,3	7,6	6,5	11,3	15,6
26	3,9	3,7	5,0	1,8	2,0	5,4	3,0	4,5	4,4	7,1	9,4	5,0	4,0	3,1	3,7	5,5	3,7	5,6	4,9	8,1	1,3	4,0	0,0	0,3	0,7	0,0	0,4	2,3	1,4	1,3	0,5	2,0	6,9	6,4	12,8	19,2
27	4,1	4,0	5,6	1,6	2,2	5,8	3,3	5,2	4,9	7,6	10,7	5,3	4,0	3,2	3,7	6,2	4,5	6,0	5,2	8,7	1,7	4,2	0,4	0,3	1,1	0,4	0,0	2,4	2,4	1,6	1,3	2,0	10,6	9,9	17,3	24,0
28	0,2	0,2	0,7	0,2	0,0	0,7	0,1	0,5	0,4	1,4	3,0	0,6	0,3	0,1	0,2	0,9	0,4	0,8	0,5	1,9	0,2	0,3	1,9	1,1	0,4	2,3	2,4	0,0	0,5	6,9	4,6	0,0	9,7	8,1	12,0	14,7
29	1,0	0,9	1,2	0,9	0,4	1,6	0,6	1,0	1,1	2,5	3,5	1,5	1,4	0,9	1,2	1,5	0,6	1,8	1,5	3,0	0,2	1,1	1,2	1,0	0,3	1,4	2,4	0,5	0,0	5,4	2,9	0,6	5,8	4,6	8,2	11,6
30	9,7	9,4	11,4	5,8	6,5	12,0	8,2	10,7	10,6	14,5	17,6	11,4	9,8	8,3	9,2	12,1	9,4	12,3	11,3	15,9	5,2	9,8	1,6	2,5	4,0	1,3	1,6	6,9	5,4	0,0	0,5	6,4	9,0	9,3	18,0	27,4
31	6,7	6,5	7,9	4,0	4,2	8,5	5,5	7,4	7,3	10,7	12,9	8,1	7,1	5,7	6,6	8,5	6,2	8,9	8,1	11,7	3,1	6,9	0,7	1,4	2,2	0,5	1,3	4,6	2,9	0,5	0,0	4,3	5,8	5,8	12,8	20,6
32	0,4	0,3	0,9	0,1	0,0	1,0	0,2	0,8	0,6	1,7	3,5	0,8	0,4	0,1	0,3	1,1	0,6	1,0	0,7	2,3	0,1	0,4	1,6	0,9	0,3	2,0	2,0	0,0	0,6	6,4	4,3	0,0	10,0	8,5	12,7	15,7
33	10,9	10,5	10,2	11,0	9,3	11,4	9,8	9,8	10,5	13,1	11,7	11,9	12,4	10,9	11,8	10,6	8,6	12,0	12,1	13,2	8,0	11,2	7,1	8,6	7,6	6,9	10,6	9,7	5,8	9,0	5,8	10,0	0,0	0,1	2,1	7,2
34	9,0	8,7	8,2	9,5	7,8	9,4	8,1	7,9	8,6	10,8	9,4	9,8	10,4	9,1	9,9	8,5	6,9	9,9	10,0	10,8	6,7	9,2	6,4	7,7	6,5	6,4	9,9	8,1	4,6	9,3	5,8	8,5	0,1	0,0	1,5	5,9
35	11,9	11,6	10,1	14,3	11,7	11,3	11,4	10,0	10,9	12,1	9,0	12,2	13,7	12,6	13,2	10,1	9,2	11,8	12,5	11,5	11,0	12,1	12,7	13,8	11,3	12,8	17,3	12,0	8,2	18,0	12,8	12,7	2,1	1,5	0,0	1,6
36	13,4	13,1	10,8	17,8	14,7	11,8	13,5	11,0	12,0	11,8	7,6	13,0	15,2	14,8	15,0	10,6	10,7	12,3	13,4	10,7	14,6	13,5	18,8	19,3	15,6	19,2	24,0	14,7	11,6	27,4	20,6	15,7	7,2	5,9	1,6	0,0

Esto es una matriz de disimilaridad.

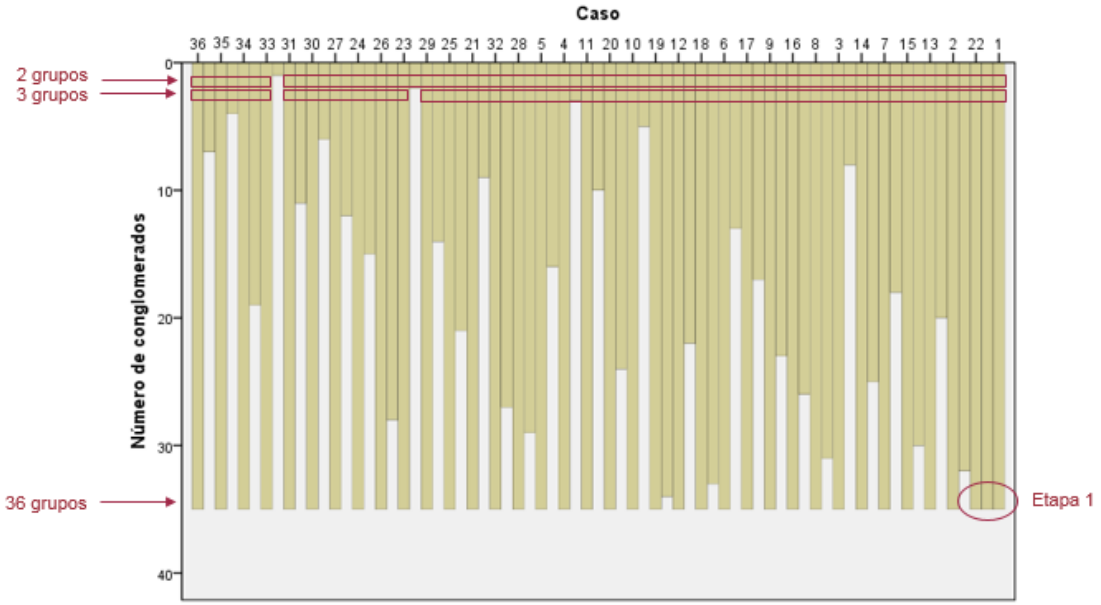
Historial de conglomeración obtenido con la clasificación por el método Ward:

[8] El archivo ACL-Alcobendas contiene la sintaxis del análisis que hemos solicitado a través del menú junto con algunos tablas descriptivas para la caracterización de los grupos y gráficos adicionales.

Historial de conglomeración

Etapa	Clúster combinado		Coeficientes	Primera aparición del clúster de etapa		Etapa siguiente
	Clúster 1	Clúster 2		Clúster 1	Clúster 2	
1	1	22	,001	0	0	4
2	12	19	,002	0	0	14
3	6	18	,006	0	0	14
4	1	2	,010	1	0	16
5	3	8	,014	0	0	10
6	13	15	,019	0	0	16
7	5	28	,024	0	0	9
8	23	26	,034	0	0	21
9	5	32	,046	7	0	20
10	3	16	,063	5	0	13
11	7	14	,079	0	0	18
12	10	20	,107	0	0	26
13	3	9	,137	10	0	19
14	6	12	,172	3	2	23
15	21	25	,210	0	0	22
16	1	13	,276	4	6	18
17	33	34	,343	0	0	32
18	1	7	,418	16	11	28
19	3	17	,498	13	0	23
20	4	5	,590	0	9	27
21	23	24	,730	8	0	24
22	21	29	,883	15	0	27
23	3	6	1,082	19	14	28
24	23	27	1,309	21	0	30
25	30	31	1,554	0	0	30
26	10	11	1,826	12	0	31
27	4	21	2,302	20	22	33
28	1	3	3,000	18	23	31
29	35	36	3,821	0	0	32
30	23	30	5,336	24	25	34
31	1	10	7,124	28	26	33
32	33	35	10,875	17	29	35
33	1	4	15,561	31	27	34
34	1	23	37,569	33	30	35
35	1	33	70,000	34	32	0

Diagrama de témpanos obtenido con el método Ward:



Dendrograma

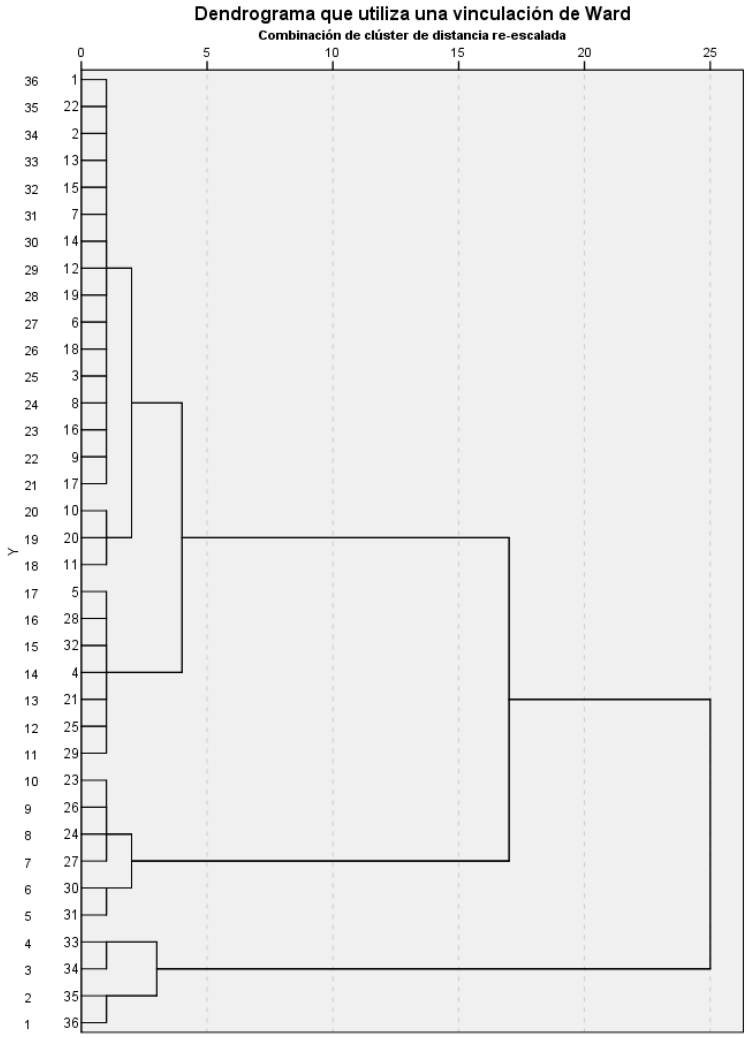


Tabla con el grupo de pertenencia:

Clúster de pertenencia

Caso	5 clústeres	4 clústeres	3 clústeres	2 clústeres
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
4	2	2	1	1
5	2	2	1	1
6	1	1	1	1
7	1	1	1	1
8	1	1	1	1
9	1	1	1	1
10	1	1	1	1
11	1	1	1	1
12	1	1	1	1
13	1	1	1	1
14	1	1	1	1
15	1	1	1	1
16	1	1	1	1
17	1	1	1	1
18	1	1	1	1
19	1	1	1	1
20	1	1	1	1
21	2	2	1	1
22	1	1	1	1
23	3	3	2	1
24	3	3	2	1
25	2	2	1	1
26	3	3	2	1
27	3	3	2	1
28	2	2	1	1
29	2	2	1	1
30	3	3	2	1
31	3	3	2	1
32	2	2	1	1
33	4	4	3	2
34	4	4	3	2
35	5	4	3	2
36	5	4	3	2

Media de las variables originales y los factores o componentes según la clasificación en 5 grupos por el método ward de las secciones censales del estudio de Alcobendas:

	CLU5_1					
	1	2	3	4	5	Total
FAC1_1	,504	,287	-,191	-2,529	-2,689	,000
FAC2_1	,552	-,380	-1,726	-,334	1,596	,000
INMIGR	33,368	28,286	26,000	9,500	8,500	28,444
PADMIM	15,474	25,429	35,667	26,000	12,000	21,167
PAROJU	15,158	12,000	9,833	2,500	3,000	12,278
PDOMEM	28,632	23,714	12,500	18,000	43,500	25,222
PDTORE	,316	,714	1,333	7,500	6,500	1,306
PEVEN	18,789	15,429	9,000	4,000	5,500	14,944
PEVENM	30,579	25,714	15,333	8,500	11,000	24,778
POPER	43,895	38,429	23,833	2,500	3,500	34,944
PPOBJO	35,737	21,000	12,833	10,500	26,000	27,111
PTECN	5,632	7,857	17,333	34,500	26,000	10,750
PTECNM	9,158	9,857	22,833	33,500	20,000	13,528
SINEST	32,053	25,286	16,667	4,500	7,500	25,278
SUPER	1,632	3,000	8,667	33,000	23,500	6,028
TACTM	23,211	26,714	33,333	32,000	27,500	26,306
TPAROM	32,632	27,571	21,333	12,000	10,000	27,361

Media de las variables originales y los factores o componentes según la clasificación en 4 grupos por el método ward de las secciones censales del estudio de Alcobendas

	CLU4_1				
	1	2	3	4	Total
FAC1_1	,504	,287	-,191	-2,609	,000
FAC2_1	,552	-,380	-1,726	,631	,000
INMIGR	33,368	28,286	26,000	9,000	28,444
PADMIM	15,474	25,429	35,667	19,000	21,167
PAROJU	15,158	12,000	9,833	2,750	12,278
PDOMEM	28,632	23,714	12,500	30,750	25,222
PDTORE	,316	,714	1,333	7,000	1,306
PEVEN	18,789	15,429	9,000	4,750	14,944
PEVENM	30,579	25,714	15,333	9,750	24,778
POPER	43,895	38,429	23,833	3,000	34,944
PPOBJO	35,737	21,000	12,833	18,250	27,111
PTECN	5,632	7,857	17,333	30,250	10,750
PTECNM	9,158	9,857	22,833	26,750	13,528
SINEST	32,053	25,286	16,667	6,000	25,278
SUPER	1,632	3,000	8,667	28,250	6,028
TACTM	23,211	26,714	33,333	29,750	26,306
TPAROM	32,632	27,571	21,333	11,000	27,361

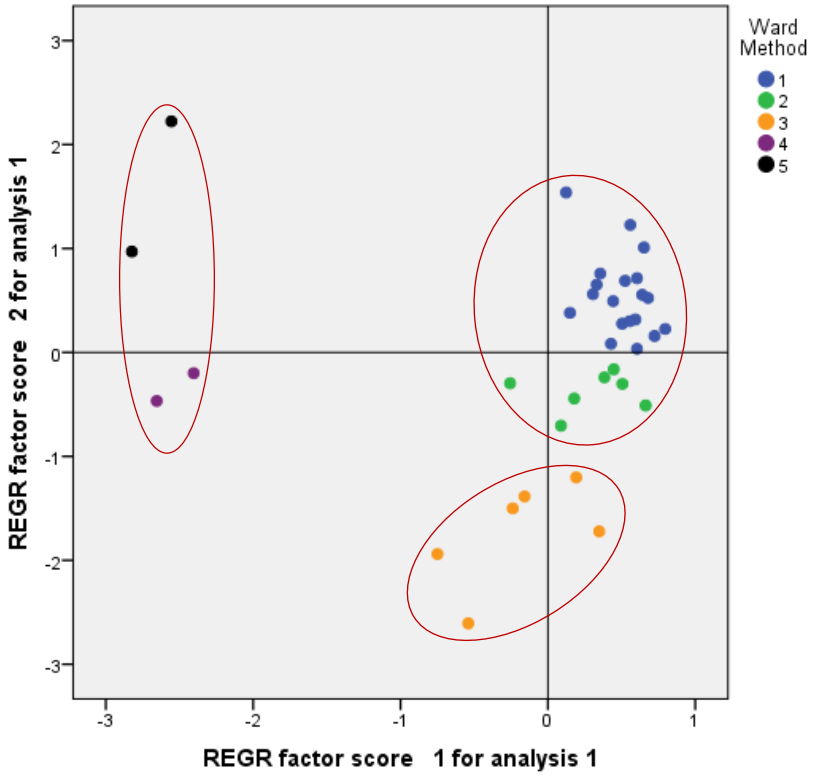
Media de las variables originales y los factores o componentes según la clasificación en 3 grupos por el método ward de las secciones censales del estudio de Alcobendas

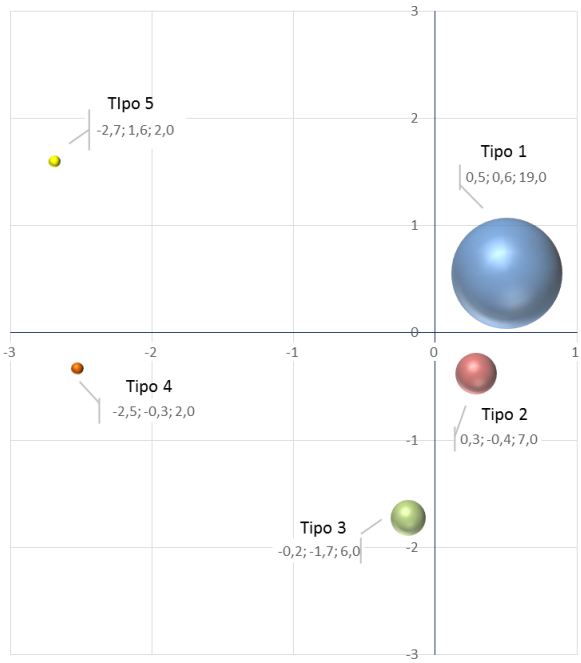
	CLU3_1			
	1	2	3	Total
FAC1_1	,445	-,191	-2,609	,000
FAC2_1	,301	-1,726	,631	,000
INMIGR	32,000	26,000	9,000	28,444
PADMIM	18,154	35,667	19,000	21,167
PAROJU	14,308	9,833	2,750	12,278
PDOMEM	27,308	12,500	30,750	25,222
PDTORE	,423	1,333	7,000	1,306
PEVEN	17,885	9,000	4,750	14,944
PEVENM	29,269	15,333	9,750	24,778
POPER	42,423	23,833	3,000	34,944
PPOBJO	31,769	12,833	18,250	27,111
PTECN	6,231	17,333	30,250	10,750
PTECNM	9,346	22,833	26,750	13,528
SINEST	30,231	16,667	6,000	25,278
SUPER	2,000	8,667	28,250	6,028
TACTM	24,154	33,333	29,750	26,306
TPAROM	31,269	21,333	11,000	27,361

Media de las variables originales y los factores o componentes según la clasificación en 2 grupos por el método ward de las secciones censales del estudio de Alcobendas

	CLU2_1		
	1	2	Total
FAC1_1	,326	-2,609	,000
FAC2_1	-,079	,631	,000
INMIGR	30,875	9,000	28,444
PADMIM	21,438	19,000	21,167
PAROJU	13,469	2,750	12,278
PDOMEM	24,531	30,750	25,222
PDTORE	,594	7,000	1,306
PEVEN	16,219	4,750	14,944
PEVENM	26,656	9,750	24,778
POPER	38,938	3,000	34,944
PPOBJO	28,219	18,250	27,111
PTECN	8,313	30,250	10,750
PTECNM	11,875	26,750	13,528
SINEST	27,688	6,000	25,278
SUPER	3,250	28,250	6,028
TACTM	25,875	29,750	26,306
TPAROM	29,406	11,000	27,361

Representación de las secciones censales, según el grupo pertenencia a la clasificación en cinco grupos por el método ward, en el espacio de los factores o componentes rotado obtenido de la ACP con los datos del estudio de Alcobendas

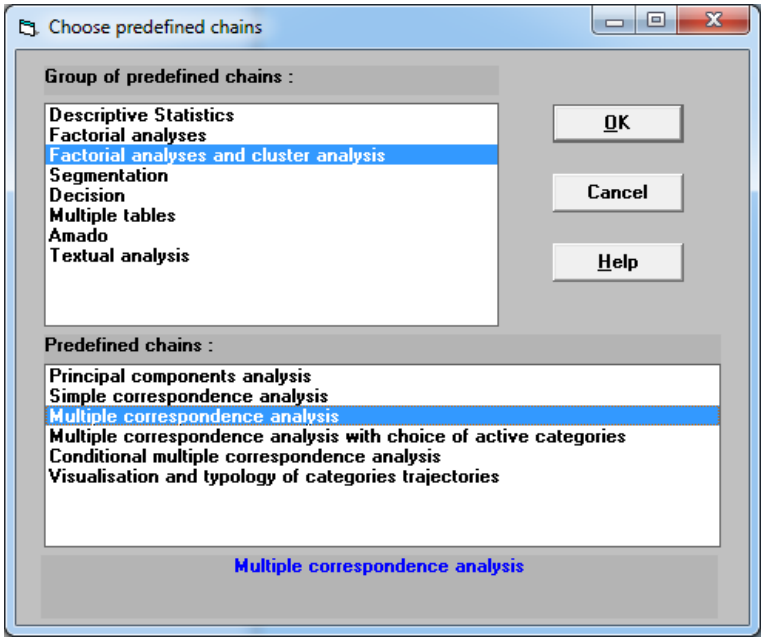




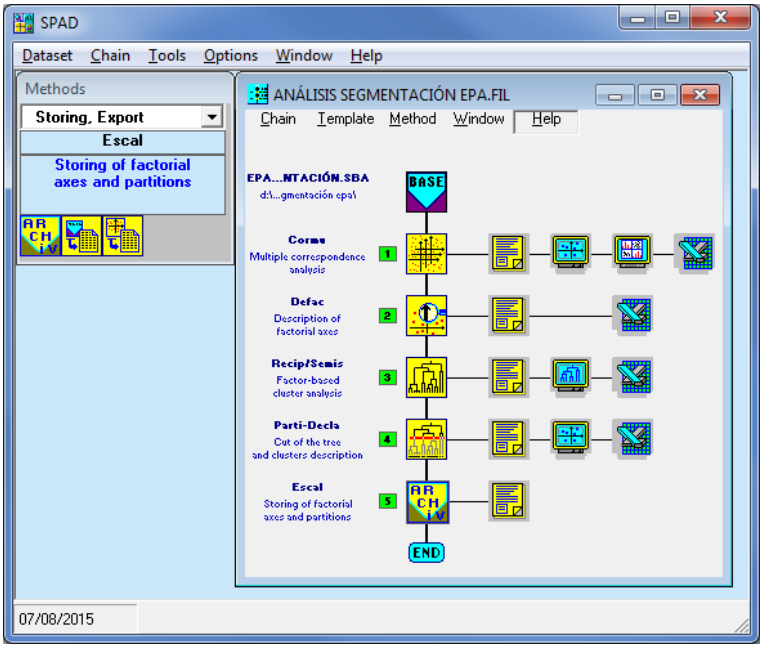
5. Análisis de clasificación con SPAD

Para ejemplificar el uso del software estadístico retomaremos el ejercicio de análisis iniciado en el capítulo anterior sobre el proceso de segmentación del mercado de trabajo de la población asalariada desde la perspectiva del empleo. Analizamos los datos de la Encuesta de Población Activa correspondientes al cuarto trimestre de 2014 a partir de un conjunto de 8 variables activas (y 40 categorías asociadas) y tres variables ilustrativas o suplementarias. Estas variables se analizaron a través de un análisis de correspondencias múltiples y se transformaron en dos factores principales de diferenciación que denominamos de forma resumida como factor de segmentación laboral al primero, con el 60% de la inercia explicada, y como factor de sector público-privado al segundo, con el 14% de la inercia. A partir de estos resultados procedemos a realizar el análisis clasificatorio.

El análisis de clasificación en SPAD se concibe conjuntamente con un análisis factorial previo en una secuencia de métodos con distintas alternativas. A través de Template / Predefined chains accedemos al cuadro de diálogo donde elegiremos el grupo de Factorial Analysis and cluster analysis. Entre las diferentes alternativas aquí veremos un análisis de correspondencias múltiples combinado con el análisis de clasificación, tal y como se muestra en la siguiente imagen del cuadro de diálogo de las cadenas predefinidas.



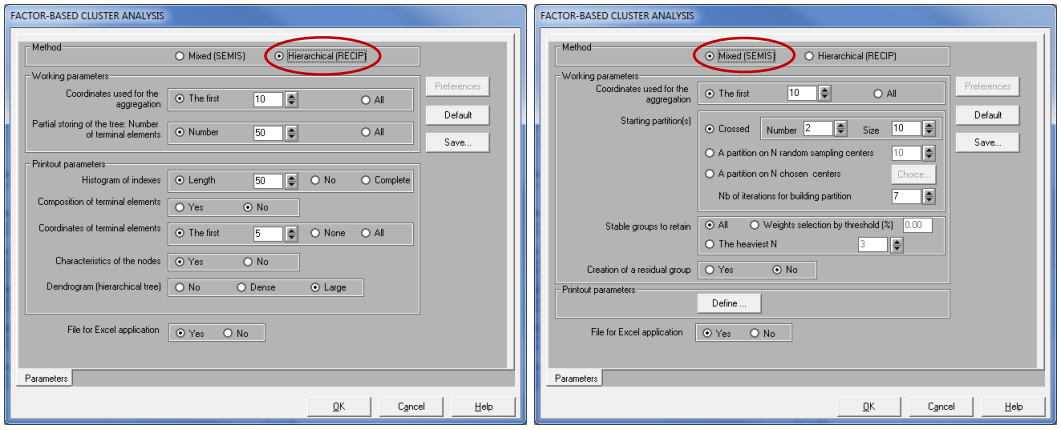
La cadena de sintaxis que se construye combina cuatro procedimientos: CORMU, DEFAC, RECIP/SEMIS y PARTI-DECLA. Adicionalmente, una vez realizado el análisis, podemos insertado el método, ESCAL, para guardar en la matriz de datos las variables factoriales y las variables tipológicas que se obtienen.



Los dos primeros procedimientos los vimos en el capítulo anterior con otro ejemplo. En este caso con los datos de la matriz EPA4T2014-Segmentación.sba seleccionaremos, desde la pestaña de Variables del procedimiento CORMU, las 8 variables V2 a V9 como activas, y las variables V10, V11 y V12 como suplementarias. En la pestaña Parameters fijaremos el número de factores o coordenadas a retener en 2 y la asignación aleatoria de categorías con baja frecuencia lo ajustaremos al 1%. Con

esta parametrización inicial ya es posible ejecutar todos los procedimientos implicados con las opciones por defecto. Comentaremos seguidamente los relativos al procedimiento de clasificación.

Con `RECIP/SEMIS` se opta por el procedimiento de clasificación jerárquico ascendente Ward (método `RECIP`) o bien por un procedimiento mixto (método `SEMIS`) que vimos anteriormente bajo la denominación de clasificación híbrida.



En el método `RECIP` método se especifica el número de factores que se considerarán, por defecto son 10. Aquí especificaremos la decisión que hayamos tomado en el análisis factorial o bien de forma automática tomará los retenidos en el ACM, en nuestro caso los dos primeros factores. A continuación vemos que se consideran por defecto 50 elementos terminales (nodos) del árbol de agregación para presentarse en los resultados.

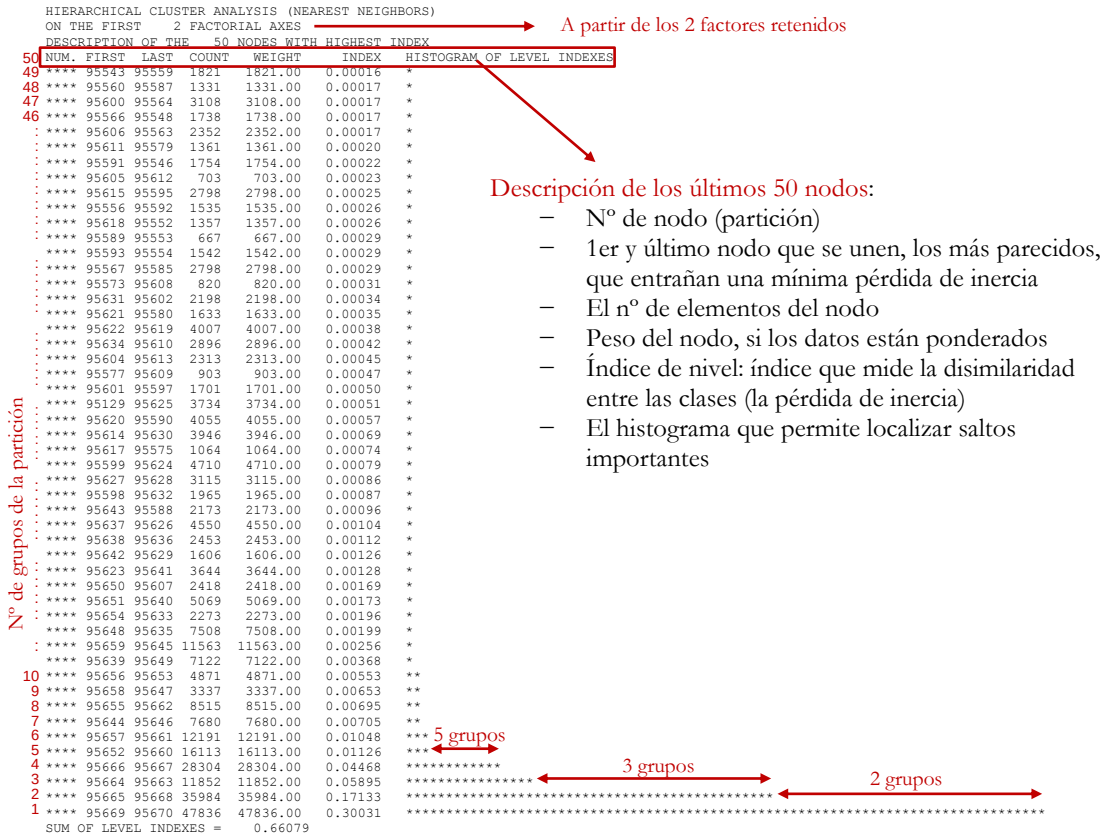
En el apartado de los parámetros de resultados se detalla en primer lugar el número de valores del histograma con las distancias a la que se forman los grupos (50 por defecto), y el listado con el grupo de pertenencia de cada unidad que sólo lo pediremos en el caso de pocas unidades y con identidad. Se extraen también coordenadas de los valores test de los elementos terminales y las características de los nodos por encima de los elementos terminales así como el dendrograma. Por defecto se procede a la exportación de los resultados a Excel. Si no tenemos una necesidad particular como regla general las opciones por defecto son válidas y no requiere que se cambien. En el análisis de segmentación hemos optado por este procedimiento.

Alternativamente el método `SEMIS` realiza una clasificación mixta y es idóneo cuando el número de casos es muy numeroso y el procedimiento jerárquico anterior no puede procesar los datos. Con este procedimiento se efectúa una primera clasificación se obtiene mediante el cruce de varias particiones de bases construidas alrededor de los centros móviles y a continuación las clases estables formadas se agregan por un método de clasificación jerárquica por el criterio de Ward. Adicionalmente se pueden optimizar las particiones obtenidas.

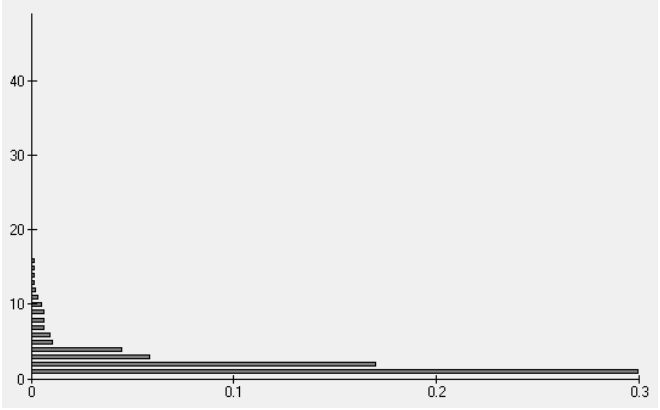
En el cuadro de diálogo se especifica en primer lugar, como antes, el número de factores y seguidamente los parámetros iniciales de partición. Por defecto opera el procedimiento de tablas cruzadas donde se especifica el número de clasificaciones que se cruzan (por defecto 2) y el tamaño o número de clases de cada clasificación (por

defecto 10, generando por tanto una tabla cruzada de 10×10). Una segunda opción es iniciar el proceso clasificatorio con un número de centros (por defecto 10) elegidos aleatoriamente, o bien en tercer lugar fijar los centros con valores concretos que se especifican. Por defecto se retienen todos los grupos estables pero existe la posibilidad de efectuar una elección así como de crear un grupo residual. Finalmente se definen los parámetros de resultados en términos del número de nodos, descripción característica y dendrograma. Por defecto se procede a la exportación de los resultados a Excel. Ante la diversidad de opciones se puede ejecutar la configuración por defecto y comparar estos resultados con alguna variación, por ejemplo, aumentando el número de clases y de particiones inicial.

Se presentan a continuación los resultados de tablas y gráficos que se generan en el análisis de segmentación con el procedimiento `RECIP`. La información que se proporciona en primer lugar es la relativa al proceso de creación de los grupos en las últimas 50 etapas con el índice creciente que se calcula en el proceso de agregación de los datos. Este valor se representa con un histograma que nos permite ver en el final del proceso clasificatorio dónde se producen los mayores saltos. En este caso podemos observar gráficamente los mayores saltos producidos que sugieren optar por clasificaciones en 2, 3 o 5 grupos.



Un histograma con mejor resolución gráfica se puede obtener desde el editor del dendrograma que comentamos después a través del menú Edit / Curve of thee level indexes que ofrece la imagen siguiente:



La tabla siguiente contiene la descripción de los 50 elementos terminales (nodos) con los valores-test de significación de cada nodo sobre cada factor retenido (valores superiores a 2 son significativos) y las coordenadas de cada nodo en los ejes factoriales. Se trata de una información detallada del proceso de formación de los grupos sin mayor interés cuando el número de unidades es elevado y formado por unidades anónimas como en este caso lo son las personas entrevistadas en la encuesta.

FACTOR SCORES AND TEST-VALUES
AXES 1 A 2

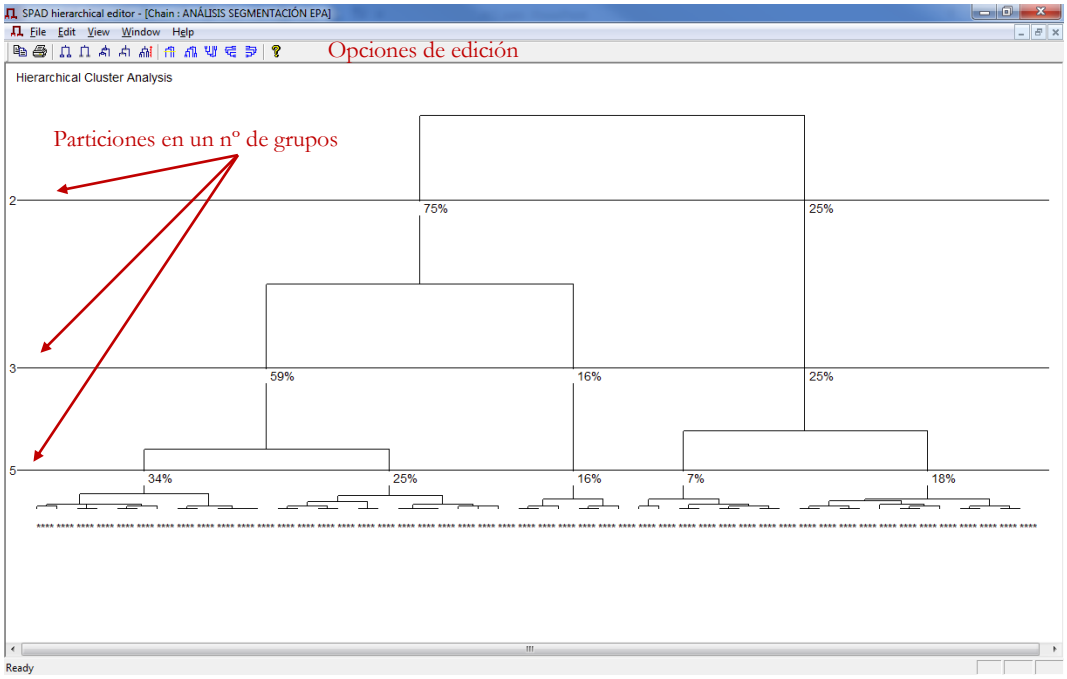
ELEMENTS				TEST-VALUES					FACTOR SCORES				
NUM	IDENT	WEIGHT	COUNT	1	2	0	0	0	1	2	0	0	0
1	****	1269.00	1269	-15.51	36.17	0.00	0.00	0.00	-0.43	1.00	0.00	0.00	0.00
2	****	2786.00	2786	-8.58	48.55	0.00	0.00	0.00	-0.16	0.89	0.00	0.00	0.00
3	****	1546.00	1546	-8.36	51.34	0.00	0.00	0.00	-0.21	1.28	0.00	0.00	0.00
4	****	1252.00	1252	-1.67	52.87	0.00	0.00	0.00	-0.05	1.47	0.00	0.00	0.00
5	****	786.00	786	2.55	22.07	0.00	0.00	0.00	0.09	0.78	0.00	0.00	0.00
6	****	2322.00	2322	6.05	49.36	0.00	0.00	0.00	0.12	1.00	0.00	0.00	0.00
7	****	1602.00	1602	6.99	52.67	0.00	0.00	0.00	0.17	1.29	0.00	0.00	0.00
8	****	585.00	585	9.94	26.59	0.00	0.00	0.00	0.41	1.09	0.00	0.00	0.00
9	****	1767.00	1767	17.44	35.73	0.00	0.00	0.00	0.41	0.83	0.00	0.00	0.00
10	****	663.00	663	14.73	13.10	0.00	0.00	0.00	0.57	0.51	0.00	0.00	0.00
11	****	1135.00	1135	23.30	28.44	0.00	0.00	0.00	0.68	0.83	0.00	0.00	0.00
12	****	400.00	400	19.38	12.69	0.00	0.00	0.00	0.97	0.63	0.00	0.00	0.00
13	****	806.00	806	-4.42	14.66	0.00	0.00	0.00	-0.15	0.51	0.00	0.00	0.00
14	****	948.00	948	-11.24	10.88	0.00	0.00	0.00	-0.36	0.35	0.00	0.00	0.00
15	****	812.00	812	3.94	12.18	0.00	0.00	0.00	0.14	0.42	0.00	0.00	0.00
16	****	549.00	549	0.87	2.91	0.00	0.00	0.00	0.04	0.12	0.00	0.00	0.00
17	****	2186.00	2186	-22.57	28.88	0.00	0.00	0.00	-0.47	0.60	0.00	0.00	0.00
18	****	1821.00	1821	-29.47	30.39	0.00	0.00	0.00	-0.68	0.70	0.00	0.00	0.00
19	****	1354.00	1354	-24.10	-1.45	0.00	0.00	0.00	-0.65	-0.04	0.00	0.00	0.00
20	****	527.00	527	-22.77	0.45	0.00	0.00	0.00	-0.99	0.02	0.00	0.00	0.00
21	****	1015.00	1015	-24.18	9.05	0.00	0.00	0.00	-0.75	0.28	0.00	0.00	0.00
22	****	472.00	472	-5.74	-6.77	0.00	0.00	0.00	-0.26	-0.31	0.00	0.00	0.00
23	****	364.00	364	-13.14	-15.68	0.00	0.00	0.00	-0.69	-0.82	0.00	0.00	0.00
24	****	1337.00	1337	-32.15	-11.72	0.00	0.00	0.00	-0.87	-0.32	0.00	0.00	0.00
25	****	1560.00	1560	-43.71	-24.29	0.00	0.00	0.00	-1.09	-0.60	0.00	0.00	0.00
26	****	1238.00	1238	-45.69	-17.58	0.00	0.00	0.00	-1.28	-0.49	0.00	0.00	0.00
27	****	1148.00	1148	-47.99	-28.51	0.00	0.00	0.00	-1.40	-0.83	0.00	0.00	0.00
28	****	165.00	165	-14.82	-18.47	0.00	0.00	0.00	-1.15	-1.44	0.00	0.00	0.00
29	****	1573.00	1573	-59.70	-48.90	0.00	0.00	0.00	-1.48	-1.21	0.00	0.00	0.00
30	****	1996.00	1996	-76.00	-64.34	0.00	0.00	0.00	-1.67	-1.41	0.00	0.00	0.00
31	****	245.00	245	3.96	-47.03	0.00	0.00	0.00	0.25	-3.00	0.00	0.00	0.00
32	****	819.00	819	-8.06	-71.19	0.00	0.00	0.00	-0.28	-2.47	0.00	0.00	0.00
33	****	302.00	302	7.07	-24.97	0.00	0.00	0.00	0.41	-1.43	0.00	0.00	0.00
34	****	365.00	365	0.12	-32.98	0.00	0.00	0.00	0.01	-1.72	0.00	0.00	0.00
35	****	416.00	416	16.67	-45.78	0.00	0.00	0.00	0.81	-2.23	0.00	0.00	0.00
36	****	287.00	287	6.87	-37.27	0.00	0.00	0.00	0.40	-2.19	0.00	0.00	0.00
37	****	562.00	562	21.10	-35.38	0.00	0.00	0.00	0.88	-1.48	0.00	0.00	0.00
38	****	341.00	341	23.74	-34.95	0.00	0.00	0.00	1.28	-1.89	0.00	0.00	0.00
39	****	398.00	398	18.29	0.91	0.00	0.00	0.00	0.91	0.05	0.00	0.00	0.00
40	****	422.00	422	27.32	4.55	0.00	0.00	0.00	1.32	0.22	0.00	0.00	0.00
41	****	627.00	627	42.86	-3.47	0.00	0.00	0.00	1.70	-0.14	0.00	0.00	0.00
42	****	1006.00	1006	44.79	-9.95	0.00	0.00	0.00	1.40	-0.31	0.00	0.00	0.00
43	****	453.00	453	16.47	-14.81	0.00	0.00	0.00	0.77	-0.69	0.00	0.00	0.00
44	****	400.00	400	27.46	-14.24	0.00	0.00	0.00	1.37	-0.71	0.00	0.00	0.00
45	****	957.00	957	40.58	-34.26	0.00	0.00	0.00	1.30	-1.10	0.00	0.00	0.00
46	****	608.00	608	44.26	-30.52	0.00	0.00	0.00	1.78	-1.23	0.00	0.00	0.00
47	****	1324.00	1324	65.16	-24.26	0.00	0.00	0.00	1.77	-0.66	0.00	0.00	0.00
48	****	989.00	989	63.66	-13.11	0.00	0.00	0.00	2.00	-0.41	0.00	0.00	0.00
49	****	811.00	811	61.39	-23.86	0.00	0.00	0.00	2.14	-0.83	0.00	0.00	0.00
50	****	520.00	520	54.96	-18.41	0.00	0.00	0.00	2.40	-0.80	0.00	0.00	0.00

Por último aparece otra tabla descriptiva de los 50 nodos de la jerarquía donde sucesivamente las columnas presenta el número del nodo y el índice de nivel del nodo, el rango de los dos nodos o grupos que se unen, los efectivos o frecuencia sin o con ponderación, y el rango del primer y último nodo terminal comprendido en el nodo estudiado. De nuevo se trata de información muy específica del proceso de agregación que nos proporciona información especialmente relevante.

DESCRIPTION OF HIERARCHY NODES
(INDEXES AS PERCENTAGES OF SUM OF INDEXES : 0.65358)

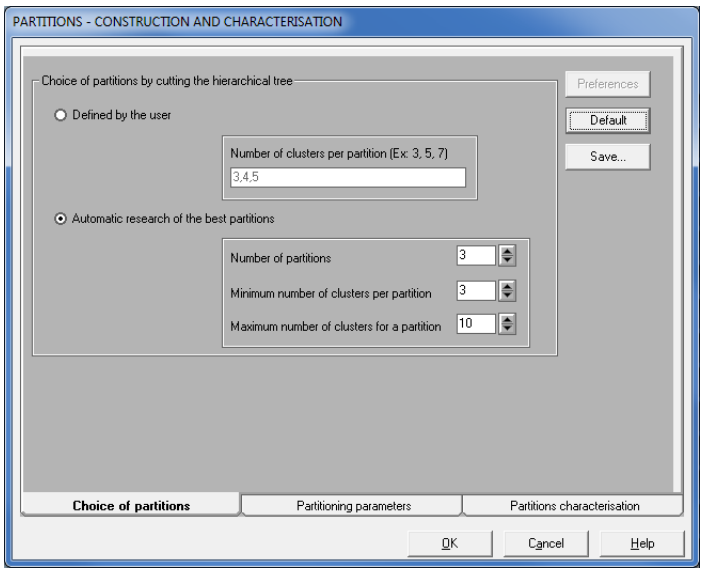
NODE		SUCCESSORS				COMPOSITION	
NUMBER	INDEX	FIRST	LAST	COUNT	WEIGHT	FIRST	LAST
51	0.03	50	49	1331	1331.00	49	50
52	0.03	6	5	3108	3108.00	5	6
53	0.03	29	28	1738	1738.00	28	29
54	0.03	9	8	2352	2352.00	8	9
55	0.03	16	15	1361	1361.00	15	16
56	0.03	14	13	1754	1754.00	13	14
57	0.03	36	35	703	703.00	35	36
58	0.04	26	25	2798	2798.00	25	26
59	0.04	12	11	1535	1535.00	11	12
60	0.04	45	44	1357	1357.00	44	45
61	0.04	34	33	667	667.00	33	34
62	0.04	21	20	1542	1542.00	20	21
63	0.04	4	3	2798	2798.00	3	4
64	0.05	40	39	820	820.00	39	40
65	0.05	59	10	2198	2198.00	10	12
66	0.05	42	41	1633	1633.00	41	42
67	0.06	18	17	4007	4007.00	17	18
68	0.06	62	19	2896	2896.00	19	21
69	0.07	48	47	2313	2313.00	47	48
70	0.07	38	37	903	903.00	37	38
71	0.08	24	23	1701	1701.00	23	24
72	0.08	30	53	3734	3734.00	28	30
73	0.09	2	1	4055	4055.00	1	2
74	0.11	27	58	3946	3946.00	25	27
75	0.11	32	31	1064	1064.00	31	32
76	0.12	7	52	4710	4710.00	5	7
77	0.13	55	56	3115	3115.00	13	16
78	0.13	46	60	1965	1965.00	44	46
79	0.15	71	22	2173	2173.00	22	24
80	0.16	65	54	4550	4550.00	8	12
81	0.17	66	64	2453	2453.00	39	42
82	0.19	70	57	1606	1606.00	35	38
83	0.20	51	69	3644	3644.00	47	50
84	0.26	78	43	2418	2418.00	43	46
85	0.27	79	68	5069	5069.00	19	24
86	0.30	82	61	2273	2273.00	33	38
87	0.30	76	63	7508	7508.00	3	7
88	0.39	87	73	11563	11563.00	1	7
89	0.56	67	77	7122	7122.00	13	18
90	0.85	84	81	4871	4871.00	39	46
91	1.00	86	75	3337	3337.00	31	38
92	1.06	83	90	8515	8515.00	39	50
93	1.08	72	74	7680	7680.00	25	30
94	1.60	85	89	12191	12191.00	13	24
95	1.72	80	88	16113	16113.00	1	12
96	6.84	94	95	28304	28304.00	1	24
97	9.02	92	91	11852	11852.00	31	50
98	26.21	93	96	35984	35984.00	1	30
99	45.95	97	98	47836	47836.00	1	50

Un último elemento informativo de este primer procedimiento es el dendrograma que permite visualizar gráficamente el proceso de agregación a través de la representación de un árbol de agregación. Se presenta en los resultados de texto y también se puede editar en una herramienta interesante incorporada por SPAD. Al clicar sobre el editor jerárquico en el icono  se visualiza el dendrograma y puede editarse para cambiar el aspecto o presentar información adicional. En particular podemos cortar el árbol de agregación a diferentes niveles simplemente clicando a la altura de cada partición. En el gráfico adjunto se han marcado las particiones en 2, 3 y 5 grupos donde se puede observar el porcentaje de casos de cada clasificación.



Todos estos resultados no nos informan todavía del contenido de los grupos o tipos que se obtienen. En general los softwares estadísticos procesan los datos con el algoritmo clasificatorio correspondiente y no incorporan dos aspectos relevantes del proceso de análisis, como vimos en particular con SPSS: la decisión del número de grupos y la descripción del contenido de los mismos. Veremos en el siguiente procedimiento de SPAD a continuación que resulta de gran ayuda para resolver ambas cuestiones.

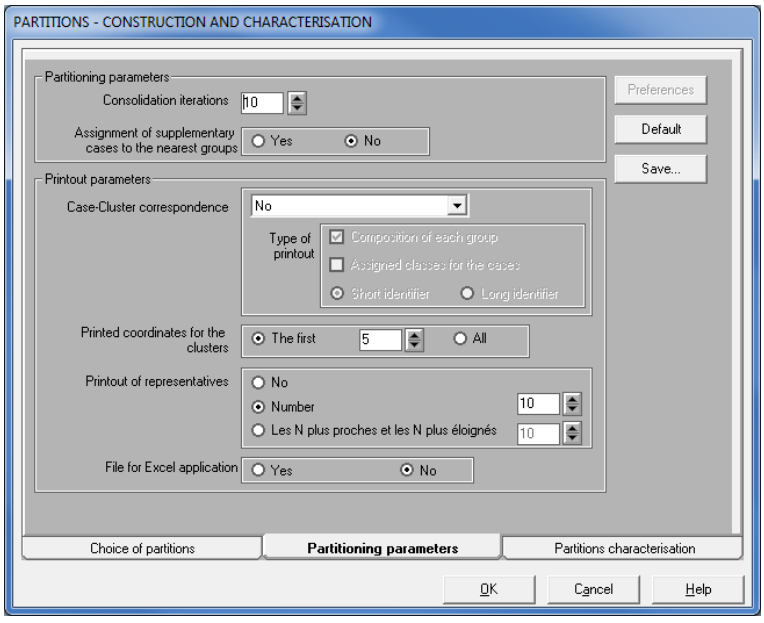
Con el procedimiento PARTI-DECLA se operan dos tareas: se realiza la partición del conjunto de unidades estableciendo los puntos de corte del dendrograma (PARTI) y se describen las clases que se obtienen (DECLA). En el primer caso el cuadro de diálogo es el siguiente con respecto a la elección de las particiones:



Por defecto se buscan las mejores particiones considerando por defecto que se extraigan las 3 mejores clasificaciones desde 3 grupos hasta 10 grupos. Este criterio se puede modificar según desee el usuario/a, por ejemplo eligiendo la mejor clasificación entre 2 y 10. En este caso hemos elegido las 3 mejores particiones entre 2 y 10. Alternativamente el usuario puede definir las particiones y fijar el número de grupos de las clasificaciones que sea analizar y luego guardar, por ejemplo si quisiéramos analizar de forma exhaustiva las tipologías que resultan entre 10 grupos y 2 relacionaríamos todos los valores enteros entre 2 y 10.

A continuación, en la siguiente pestaña, se detallan los parámetros de construcción de las particiones donde es posible realizar una consolidación u optimización de la clasificación que se obtiene con el método Ward a través de varias iteraciones con el método de centros móviles y con el objetivo de mejorar la homogeneidad de las clases y así aumentar la inercia o varianza explicada (entre-grupos) por la partición. Los centros iniciales son los obtenidos en cada partición que genera el método Ward. Si se considera un valor 0 no habrá consolidación y los resultados de la corte del árbol de agregación se guardan directamente. Si es mayor a 0, se producirá la consolidación, con el número máximo de iteraciones que se consideran (por defecto 10). Los cálculos se detienen cuando el aumento de la inercia de entre-grupos se reduce de una iteración a la siguiente. Si el valor de iteraciones es positivo se extraen las asignaciones de los casos después de la consolidación, si es negativo se extraen antes y después.

Para la presentación de resultados se puede especificar el listado de casos de cada grupo, las coordenadas de las clases o grupos, los casos representativos de cada partición (los parangones), 10 por defecto, que son los más próximos al centro de cada grupo (también se pueden elegir los más alejados). Finalmente disponemos de la opción que controla la creación de un fichero Excel con los resultados.



Los resultados del análisis de segmentación generan tres particiones entre 2 y 10 como las mejores opciones: las de 3, 5 y 7 clases o tipos, como se puede apreciar en la tabla siguiente. Si solicitamos la mejor opción de ellas la mejor solución es la de 3 tipos.

Daremos cuenta a continuación de los resultados destacando esta mejor y más parsimoniosa clasificación.

```
BUILDING UP PARTITIONS
DETERMINING THE BEST PARTITIONS
RESEARCH OF IRREGULARITIES
+--------------+--------------+------------------------------------------------------+
| IRREGULARITY | IRREGULARITY |                                                      |
|    BETWEEN   |    VALUE     |                                                      |
+--------------+--------------+------------------------------------------------------+
| 95669-- 95670|      -117.27 | ***************************************************  |
| 95667-- 95668|       -35.30 | ****************                                     |
| 95665-- 95666|        -3.65 | **                                                   |
+--------------+--------------+------------------------------------------------------+
LIST OF THE BEST   3 PARTITION BETWEEN 2 AND 10 CLUSTERS
  1 - PARTITION IN   3 CLUSTERS
  2 - PARTITION IN   5 CLUSTERS
  3 - PARTITION IN   7 CLUSTERS
```

Se presenta en primer lugar la distribución de frecuencias y los nodos, de los últimos 50, que pertenecen a cada grupo:

```
CUT "a" OF THE TREE INTO   3 CLUSTERS
CLUSTERS FORMATION (ON ACTIVE CASES)
SUMMARY DESCRIPTION
+---------+----------+-------------+------------+
| CLUSTER |  COUNT   |   WEIGHT    |   CONTENT  |
+---------+----------+-------------+------------+
|   aa1a  |  28304   | 28304.00    |   1 TO  24 |
|   aa2a  |   7680   |  7680.00    |  25 TO  30 |
|   aa3a  |  11852   | 11852.00    |  31 TO  50 |
+---------+----------+-------------+------------+
```

Las coordenadas y valores-test de los tres tipos obtenidos se pueden analizar antes y después de la consolidación u optimización (la aplicación del método de centros móviles a la clasificación obtenida por el método Ward). En particular vemos el cambio de efectivos como resultado de la redistribución de la clasificación inicial: la clase 1 del 59% al 52%, la clase 2 del 16% al 26% y la clase 3 del 25% al 22%.

```
LOADINGS AND TEST-VALUES BEFORE CONSOLIDATION
AXES  1 A  2
+-----------------------------------------+-------------------------+-----------------------------------+----------+
|                CLUSTERS                 |       TEST-VALUES       |             LOADINGS              |          |
|-----------------------------------------|-------------------------|-----------------------------------|----------|
| IDEN - LABEL               COUNT  ABS.WT. |   1     2     0     0     0 |    1      2      0      0      0 |  DISTO.  |
+-----------------------------------------+-------------------------+-----------------------------------+----------+
|   CUT "a" OF THE TREE INTO   3 CLUSTERS |                         |                                   |          |
|                                         |                         |                                   |          |
| aa1a - CLUSTER  1 /  3     28304  28304.00 | -42.1 100.0   0.0   0.0   0.0 |  -0.10   0.36   0.00   0.00   0.00 |     0.14 |
| aa2a - CLUSTER  2 /  3      7680   7680.00 |-100.0 -93.0   0.0   0.0   0.0 |  -0.86  -0.52   0.00   0.00   0.00 |     1.01 |
| aa3a - CLUSTER  3 /  3     11852  11852.00 | 100.0-100.0   0.0   0.0   0.0 |   0.79  -0.54   0.00   0.00   0.00 |     0.91 |
+-----------------------------------------+-------------------------+-----------------------------------+----------+

CLUSTERING CONSOLIDATION
AROUND CENTERS OF THE   3 CLUSTERS ACHIEVED BY 10 ITERATIONS WITH MOVING CENTERS
BETWEEN-CLUSTERS INERTIA INCREASE
+-----------+---------------+---------------+---------------+
| ITERATION | TOTAL INERTIA | INTER-CLUSTERS|     RATIO     |
|           |               |     INERTIA   |               |
+-----------+---------------+---------------+---------------+
|     0     |       0.66110 |       0.47164 |       0.71343 |
|     1     |       0.66110 |       0.49848 |       0.75401 |
|     2     |       0.66110 |       0.50121 |       0.75814 |
|     3     |       0.66110 |       0.50171 |       0.75890 |
|     4     |       0.66110 |       0.50176 |       0.75897 |
|     5     |       0.66110 |       0.50177 |       0.75899 |
+-----------+---------------+---------------+---------------+
STOP AFTER ITERATION  5. RELATIVE INCREASE OF BETWEEN-CLUSTER INERTIA
WITH RESPECT TO THE PREVIOUS ITERATION IS ONLY 0.003 %.
INERTIA DECOMPOSITION
COMPUTED ON   2 AXES.
+-----------------+-----------------+------------+-------------------+-----------------+
|                 |    INERTIAS     |   COUNTS   |      WEIGHTS      |    DISTANCES    |
| INERTIAS        | BEFORE   AFTER  |BEFORE AFTER|  BEFORE    AFTER  | BEFORE   AFTER  |
+-----------------+-----------------+------------+-------------------+-----------------+
|                 |                 |            |                   |                 |
| BETWEEN CLUSTERS | 0.4716   0.5018 |            |                   |                 |
|                 |                 |            |                   |                 |
| WITHIN CLUSTER  |                 |            |                   |                 |
|                 |                 |            |                   |                 |
| CLUSTER   1 /  3 | 0.0874   0.0542 |28304  25067 | 28304.00 25067.00 | 0.1426   0.1868 |
| CLUSTER   2 /  3 | 0.0092   0.0535 | 7680  12284 |  7680.00 12284.00 | 1.0050   0.6989 |
| CLUSTER   3 /  3 | 0.0928   0.0517 |11852  10485 | 11852.00 10485.00 | 0.9118   1.0240 |
|                 |                 |            |                   |                 |
| TOTAL INERTIA   | 0.6611   0.6611 |            |                   |                 |
+-----------------+-----------------+------------+-------------------+-----------------+
RATIO INTER INERTIA / TOTAL INERTIA) : BEFORE .. 0.7134
                                      AFTER .. 0.7590
```

```
LOADINGS AND TEST-VALUES AFTER CONSOLIDATION
AXES  1 A  2
+-------------------------------------------+-----------------------------+-----------------------------------+----------+
|                CLUSTERS                   |          TEST-VALUES        |              LOADINGS             |          |
|-------------------------------------------|-----------------------------|-----------------------------------|----------|
| IDEN - LABEL              COUNT  ABS.WT.  |   1     2     0     0     0 |   1     2     0     0     0   | DISTO.   |
+-------------------------------------------+-----------------------------+-----------------------------------+----------+
|   CUT "a" OF THE TREE INTO   3 CLUSTERS                                                                    |
|                                                                                                            |
| aa1a - CLUSTER  1 /  3     25067  25067.00 | -17.0 100.0   0.0   0.0   0.0 | -0.05  0.43  0.00  0.00  0.00 |     0.19 |
| aa2a - CLUSTER  2 /  3     12284  12284.00 |-100.0-100.0   0.0   0.0   0.0 | -0.68 -0.49  0.00  0.00  0.00 |     0.70 |
| aa3a - CLUSTER  3 /  3     10485  10485.00 | 100.0 -99.0   0.0   0.0   0.0 |  0.90 -0.46  0.00  0.00  0.00 |     1.02 |
+-------------------------------------------+-----------------------------+-----------------------------------+----------+
```

La información de las distancia al centro o la media (DISTO o DISTANCES) así como las coordenadas de cada clase, que definen el centro de cada grupo en el espacio factorial, nos permiten situar la ubicación de cada tipo en el espacio social que emergió del análisis de correspondencias. Más tarde veremos la representación gráfica de la tipología resultante y acabaremos de interpretar los perfiles definitorios de sus contenidos.

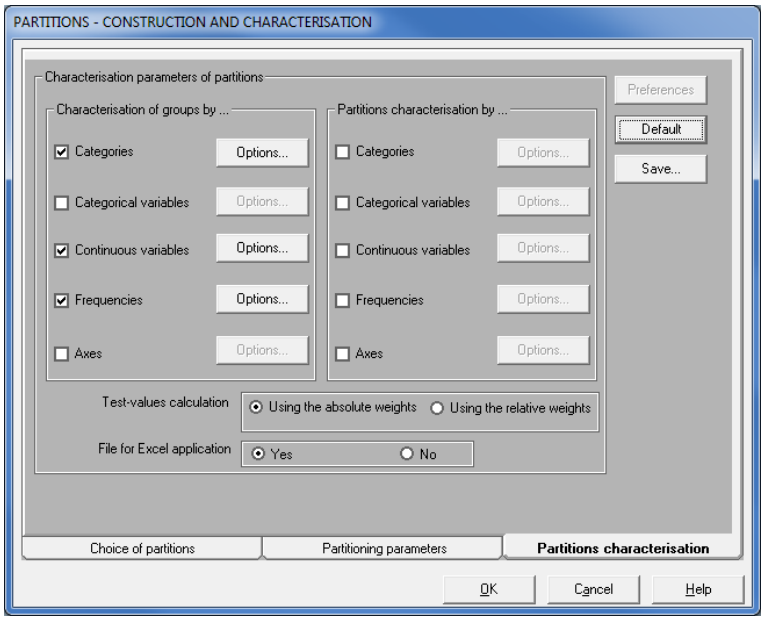
En este proceso de cambio resultado de la consolidación se presenta una tabla intermedia que relata la variación experimentada en la varianza explicada por la partición en cada momento del algoritmo iterativo (columna RATIO). Podemos observar la mejora progresiva a lo largo de las 5 iteraciones que en este caso han sido necesarias para estabilizar y alcanzar al resultado final, pasando la varianza o inercia explicada del 71,3% al 75,9%, hecho que supone una mejora de la homogeneidad interna de los grupos.

La tabla siguiente que aparece en los resultados del procedimiento PARTI es la relativa a los individuos característicos o parangones, es decir, los individuos más parecidos (próximos en el espacio) a media del grupo (o centro de gravedad).

```
CLUSTERS REPRESENTATIVES
CLUSTER   1/  3
COUNT: ****
-------------------------------------------------------------------------------
|RK | DISTANCE  | IDENT.              ||RK | DISTANCE  | IDENT.              |
+---+-----------+---------------------++---+-----------+---------------------+
|  1|    0.00009|Case n° 106          ||  2|    0.00009|Case n° 896          |
|  3|    0.00009|Case n° 1199         ||  4|    0.00009|Case n° 3546         |
|  5|    0.00009|Case n° 2679         ||  6|    0.00009|Case n° 45101        |
|  7|    0.00009|Case n° 43187        ||  8|    0.00009|Case n° 42329        |
|  9|    0.00009|Case n° 41948        || 10|    0.00009|Case n° 41281        |
+---+-----------+---------------------++---+-----------+---------------------+
CLUSTER   2/  3
COUNT: ****
-------------------------------------------------------------------------------
|RK | DISTANCE  | IDENT.              ||RK | DISTANCE  | IDENT.              |
+---+-----------+---------------------++---+-----------+---------------------+
|  1|    0.00089|Case n° 42983        ||  2|    0.00157|Case n° 6810         |
|  3|    0.00157|Case n° 9020         ||  4|    0.00200|Case n° 5407         |
|  5|    0.00200|Case n° 26374        ||  6|    0.00200|Case n° 9064         |
|  7|    0.00200|Case n° 41291        ||  8|    0.00200|Case n° 41290        |
|  9|    0.00200|Case n° 38312        || 10|    0.00216|Case n° 46870        |
+---+-----------+---------------------++---+-----------+---------------------+
CLUSTER   3/  3
COUNT: ****
-------------------------------------------------------------------------------
|RK | DISTANCE  | IDENT.              ||RK | DISTANCE  | IDENT.              |
+---+-----------+---------------------++---+-----------+---------------------+
|  1|    0.00029|Case n° 35152        ||  2|    0.00046|Case n° 31316        |
|  3|    0.00064|Case n° 887          ||  4|    0.00064|Case n° 27889        |
|  5|    0.00064|Case n° 28146        ||  6|    0.00066|Case n° 21424        |
|  7|    0.00066|Case n° 25201        ||  8|    0.00074|Case n° 32972        |
|  9|    0.00079|Case n° 24410        || 10|    0.00079|Case n° 33819        |
+---+-----------+---------------------++---+-----------+---------------------+
```

Estos individuos pueden ser estudiados con detenimiento a modo de "tipo ideal" del grupo, o incluso definir el perfil para la selección de las características de las personas que quieran ser entrevistadas en un estudio cualitativo complementario en el marco de en un diseño de investigación secuencial donde primero se analizan los datos cuantitativos y, en función de éstos, los cualitativos.

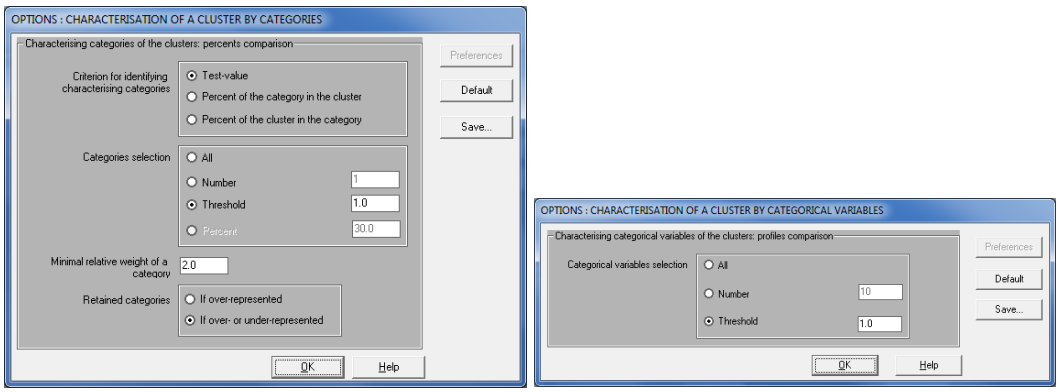
El segundo aspecto del procedimiento es la caracterización de las particiones que nos permite especificar qué información se presenta para dar cuenta, de forma descriptiva y a partir de pruebas de significación estadística, del perfil característico y definitorio de cada clase o tipo obtenido. Las opciones por defecto habitualmente son suficientes pero pueden configurarse a partir de dos posibilidades de caracterización: para cada grupo o clase de la partición o bien para el conjunto de la partición[9]. Todos los elementos utilizados en el análisis (variables cuantitativas y cualitativas, ya sean activas o suplementarias, y las variables factoriales) se relacionan con la clasificación obtenida y son los elementos caracterizantes. Los elementos característicos se presentan por orden de importancia según el criterio estadístico derivado de un valor-test al que se asocia una probabilidad (calculado bien con pesos absolutos, opción por defecto, o relativos).



La primera opción es la caracterización de cada clase de la clasificación obtenida. Se realiza con la ayuda del estadístico chi-cuadrado comparando la categoría en cuestión en la clase con respecto al total de casos. Se trata de ver si los porcentajes son iguales o bien existe una sobrerrepresentación o una infrarrepresentación de la categoría en cuestión en el grupo analizado. Con el botón de opciones podemos precisar, en primer lugar, el criterio de ordenación de las categorías características; por defecto, se realiza según el valor-test, aunque se puede cambiar al porcentaje de una categoría en la clase o al porcentaje de la clase en el grupo. En segundo lugar, las categorías que se presentan son aquellas que tienen un valor superior a 2 en el valor-test, es decir, una probabilidad asociada del 1% (o 0,01); esta opción por defecto se puede modificar al total de categorías, a un número dado o a un determinado porcentaje. En tercer lugar se define el porcentaje mínimo que debe tener una categoría caracterizante para formar parte del procedimiento, por defecto, el 2%. Finalmente se puede optar por presentar tanto las categorías sobrerrepresentadas como las infrarrepresentadas o bien solamente las primeras.

[9] El procedimiento es equivalente a DEMOD, de descripción de modalidades o categorías de una variable, que tuvimos ocasión de comentar en el capítulo III.1.

El resto de las opciones permiten caracterizar las clases o grupos a partir de variables categóricas, continuas, en forma de frecuencia o por los ejes factoriales. En todos los casos el botón de opciones abre un cuadro de diálogo similar donde se detalla el criterio de selección de los elementos caracterizantes. Como anteriormente, por defecto, las categorías que se presentan son aquellas que tienen una probabilidad asociada del 1%, si bien se puede modificar para elegir todos los elementos o bien un número dado.



Las mismas alternativas de caracterización se pueden plantear en relación al conjunto de cada partición. Por defecto no se solicitan y en cada caso se puede especificar, de forma similar a lo que hemos visto, los elementos que se seleccionan.

Con el botón Save que aparece en los distintos cuadros de diálogo que hemos visto se puede almacenar la configuración por defecto deseada y recuperarla en cualquier momento con Default.

En las tablas que siguen se presenta la caracterización de la tipología de 3 tipos de segmentos de empleo. Para cada grupo se presentan primero las categorías características que suponen un porcentaje de casos sobrerrepresentado respecto de la media de la muestra, ordenados por importancia según el valor-test que aparece en la primera columna que se acompaña de la probabilidad asociada. A continuación se disponen las categorías infrarrepresentadas, las menos características del grupo, con signo negativo en el valor-test.

La información de las tablas se completa con los porcentajes de las distintas categorías características de cada variable en el grupo, ya sea activa o ilustrativa, con dos columnas finales referidas al identificador de la categoría y la frecuencia de la categoría en cuestión en el conjunto de la muestra.

```
DESCRIPTION OF: CUT "a" OF THE TREE INTO   3 CLUSTER
CLUSTERS CHARACTERISATION BY CATEGORIES
CHARACTERISATION BY CATEGORIES OF CLUSTERS OR CATEGORIES
OF CUT "a" OF THE TREE INTO   3 CLUSTERS

CLUSTER  1 /  3
```

T.VALUE	PROB.	GRP/CAT	CAT/GRP	GLOBAL	CHARACTERISTIC CATEGORIES	OF VARIABLES	IDEN	WEIGHT
				52.40	**CLUSTER 1 / 3**		**aa1a**	**25067**
145.15	0.000	69.13	99.90	75.73	Indefinido	Duración del contrato	DU01	36228
145.15	0.000	69.13	99.90	75.73	Indefinido	Tipo de contrato	CO01	36228
129.44	0.000	67.62	98.37	76.24	Privada	Tipo de empresa asalariados	EM02	36468
49.75	0.000	74.04	29.64	20.98	Comercio	Sector de actividad	AC06	10034
40.28	0.000	67.08	35.66	27.86	4-10 años	Tiempo en la empresa	AN03	13326
36.91	0.000	65.90	34.91	27.76	Secundaria 1a	Nivel de estudios	ED03	13279
35.00	0.000	67.01	29.04	22.71	11-20 años	Tiempo en la empresa	AN04	10862
32.79	0.000	80.96	9.24	5.98	Industria 2	Sector de actividad	AC03	2862
32.52	0.000	76.80	11.81	8.06	Operadores	Ocupación	OC08	3854
28.98	0.000	70.33	15.68	11.68	Financiero-Profesion	Sector de actividad	AC08	5588
27.39	0.000	78.12	7.85	5.26	Industria 1	Sector de actividad	AC02	2518
25.42	0.000	69.05	13.96	10.60	Cualificado industri	Ocupación	OC07	5069
24.89	0.000	78.17	6.51	4.37	Industria 3	Sector de actividad	AC04	2089
17.64	0.000	66.29	9.64	7.62	Otros servicios	Sector de actividad	AC10	3646
17.38	0.000	56.30	54.70	50.92	Varón	Sexo	SE01	24356
17.13	0.000	76.26	3.61	2.48	Directivos	Ocupación	OC01	1188
14.58	0.000	66.18	6.85	5.43	Primaria	Nivel de estudios	ED02	2596
14.58	0.000	61.52	13.71	11.68	Administrativos	Ocupación	OC04	5587
13.95	0.000	64.06	8.39	6.86	Transporte-Comunicac	Sector de actividad	AC07	3283
13.86	0.000	61.54	12.47	10.62	Técnicos apoyo	Ocupación	OC03	5081
12.22	0.000	60.75	11.57	9.98	Secundaria 2a Profes	Nivel de estudios	ED05	4775
9.86	0.000	57.71	16.75	15.21	40 a 44 años	Edad	ED09	7274
9.20	0.000	56.36	23.66	22.00	Trabajador servicios	Ocupación	OC05	10524
8.43	0.000	57.40	14.14	12.91	Secundaria 2a Genera	Nivel de estudios	ED04	6176
8.39	0.000	53.29	83.91	82.52	Completa	Tipo de jornada	JO01	39472
8.10	0.000	60.63	5.54	4.79	Construcción	Sector de actividad	AC05	2291
6.78	0.000	56.08	16.16	15.10	35 a 39 años	Edad	ED08	7224
6.70	0.000	56.07	15.84	14.80	Elementales	Ocupación	OC09	7080
6.25	0.000	56.49	11.72	10.87	2-3 años	Tiempo en la empresa	AN02	5199
5.98	0.000	55.62	16.23	15.29	45 a 49 años	Edad	ED10	7314
5.94	0.000	61.77	2.42	2.05	Española y doble nac	Nacionalidad	NA02	981
4.60	0.000	56.71	6.08	5.62	Extranjera	Nacionalidad	NA03	2689
2.97	0.001	54.34	11.38	10.97	30 a 34 años	Edad	ED07	5250
-7.15	0.000	51.93	91.50	92.33	Española	Nacionalidad	NA01	44166
-8.39	0.000	48.23	16.09	17.48	Parcial	Tipo de jornada	JO02	8364
-9.29	0.000	40.44	2.36	3.06	Primario	Sector de actividad	AC01	1464
-12.38	0.000	43.31	7.31	8.85	25 a 29 años	Edad	ED06	4232
-15.78	0.000	45.54	18.83	21.67	Más de 20 años	Tiempo en la empresa	AN05	10366
-17.38	0.000	48.36	45.30	49.08	Mujer	Sexo	SE02	23480
-23.77	0.000	27.04	2.21	4.28	20 a 24 años	Edad	ED05	2045
-54.21	0.000	38.14	31.12	42.76	Superior	Nivel de estudios	ED06	20453
-63.39	0.000	0.00	0.00	5.44	Más de 6 meses	Duración del contrato	DU04	2602
-67.35	0.000	0.00	0.00	6.11	Hasta 6 meses	Duración del contrato	DU03	2921
-77.17	0.000	14.73	4.75	16.90	Hasta 1 año	Tiempo en la empresa	AN01	8083
-90.19	0.000	10.52	3.61	17.96	Profesionales	Ocupación	OC02	8590
-95.10	0.000	0.14	0.03	11.75	Hasta 1 mes	Duración del contrato	DU02	5622
-99.99	0.000	0.00	0.00	0.00	5 a 9 años	Edad	ED02	0
-99.99	0.000	0.00	0.00	0.00	0 a 4 años	Edad	ED01	0
-99.99	0.000	0.00	0.00	0.00	10 a 15 años	Edad	ED03	0
-99.99	0.000	0.00	0.00	0.00	Sin datos	Tipo de empresa asalariados	EM04	0
-99.99	0.000	0.00	0.00	0.00	Sin datos	Tiempo en la empresa	AN06	0
-99.99	0.000	0.00	0.00	0.00	No asalariado	Tipo de empresa asalariados	EM03	0
-99.99	0.000	0.00	0.00	0.00	65 o más años	Edad	ED14	0
-99.99	0.000	0.00	0.00	0.00	Sin datos	Sector de actividad	AC11	0
-99.99	0.000	0.00	0.00	0.00	Sin datos	Tipo de jornada	JO03	0
-99.99	0.000	0.00	0.00	0.00	Sin datos	Nivel de estudios	ED07	0
-99.99	0.000	0.00	0.00	0.00	Sin datos	Tipo de contrato	CO03	0
******	0.000	9.17	5.15	29.39	Administración	Sector de actividad	AC09	14061
******	0.000	3.59	1.63	23.76	Pública	Tipo de empresa asalariados	EM01	11368
******	0.000	0.21	0.10	24.27	Temporal	Tipo de contrato	CO02	11608

Cada característica se expresa en tres porcentajes. GRP/CAT es el tanto por ciento de cada clase del grupo sobre el total de la categoría; CAT/GRP es el porcentaje de la categoría sobre el total de la clase; y GLOBAL es el porcentaje de la categoría para el conjunto de los datos. Es decir, si colocamos la variable clasificatoria en filas y la variable de la categoría en las columnas se trata de los porcentajes en columna, en fila, y el marginal de columna, respectivamente. Así, por ejemplo, si consideramos la primera categoría característica de la clase 1, contrato indefinido de la variable duración del contrato, observamos como el 69,13% de los contratados de forma indefinida son de este primer grupo, el 99,9% del grupo tiene contrato indefinido, mientras que en el conjunto de la muestra tiene contrato indefinido en el 75,73% de los casos, tal y como se destaca en la tabla de contingencia que se presenta a continuación.

Casos % fila % columna	Duración del contrato					
	Indefinido	Hasta 1 mes	Hasta 6 meses	Más de 6 meses	Sin datos	Total
Cluster 1/3	25043 **99,90** **69,13**	8 0,03 0,14	0 0,00 0,00	0 0,00 0,00	16 0,06 3,46	25067 100,00 52,40
Cluster 2/3	10999 89,54 30,36	575 4,68 10,23	85 0,69 2,91	574 4,67 22,06	51 0,42 11,02	12284 100,00 25,68
Cluster 3/3	186 1,77 0,51	5039 48,06 89,63	2836 27,05 97,09	2028 19,34 77,94	396 3,78 85,53	10485 100,00 21,92
Total	36228 **75,73** 100,00	5622 11,75 100,00	2921 6,11 100,00	2602 5,44 100,00	463 0,97 100,00	47836 100,00 100,00

CLUSTER 2 / 3

T.VALUE	PROB.	PERCENTAGES GRP/CAT	PERCENTAGES CAT/GRP	PERCENTAGES GLOBAL	CHARACTERISTIC CATEGORIES	OF VARIABLES	IDEN	WEIGHT
				25.68	**CLUSTER 2 / 3**		**aa2a**	**12284**
167.65	0.000	77.64	88.87	29.39	Administración	Sector de actividad	AC09	14061
167.03	0.000	86.86	80.38	23.76	Pública	Tipo de empresa asalariados	EM01	11368
114.48	0.000	77.56	54.23	17.96	Profesionales	Ocupación	OC02	8590
88.93	0.000	46.04	76.65	42.76	Superior	Nivel de estudios	ED06	20453
69.49	0.000	53.37	45.03	21.67	Más de 20 años	Tiempo en la empresa	AN05	10366
45.19	0.000	29.46	94.65	82.52	Completa	Tipo de jornada	JO01	39472
44.07	0.000	30.36	89.54	75.73	Indefinido	Tipo de contrato	CO01	36228
44.07	0.000	30.36	89.54	75.73	Indefinido	Duración del contrato	DU01	36228
31.81	0.000	27.30	98.14	92.33	Española	Nacionalidad	NA01	44166
21.02	0.000	29.95	57.25	49.08	Mujer	Sexo	SE02	23480
20.57	0.000	38.17	15.54	10.45	55 a 59 años	Edad	ED12	5001
17.78	0.000	34.67	19.10	14.14	50 a 54 años	Edad	ED11	6766
15.15	0.000	39.14	7.97	5.23	60 a 64 años	Edad	ED13	2501
10.83	0.000	29.72	26.28	22.71	11-20 años	Tiempo en la empresa	AN04	10862
7.20	0.000	29.11	17.33	15.29	45 a 49 años	Edad	ED10	7314
-3.81	0.000	20.96	2.03	2.48	Directivos	Ocupación	OC01	1188
-4.39	0.000	22.06	4.67	5.44	Más de 6 meses	Duración del contrato	DU04	2602
-4.95	0.000	22.83	9.44	10.62	Técnicos apoyo	Ocupación	OC03	5081
-5.95	0.000	22.62	11.37	12.91	Secundaria 2a Genera	Nivel de estudios	ED04	6176
-7.42	0.000	22.20	13.06	15.10	35 a 39 años	Edad	ED08	7224
-11.70	0.000	10.81	0.86	2.05	Española y doble nac	Nacionalidad	NA02	981
-12.03	0.000	21.85	23.71	27.86	4-10 años	Tiempo en la empresa	AN03	13326
-12.51	0.000	18.76	8.02	10.97	30 a 34 años	Edad	ED07	5250
-14.00	0.000	15.84	4.23	6.86	Transporte-Comunicac	Sector de actividad	AC07	3283
-19.87	0.000	18.40	15.76	22.00	Trabajador servicios	Ocupación	OC05	10524
-21.02	0.000	21.56	42.75	50.92	Varón	Sexo	SE01	24356
-22.94	0.000	12.77	4.97	9.98	Secundaria 2a Profes	Nivel de estudios	ED05	4775
-25.14	0.000	10.80	3.72	8.85	25 a 29 años	Edad	ED06	4232
-27.81	0.000	0.75	0.09	3.06	Primario	Sector de actividad	AC01	1464
-27.98	0.000	3.33	0.55	4.28	20 a 24 años	Edad	ED05	2045
-29.70	0.000	2.54	0.43	4.37	Industria 3	Sector de actividad	AC04	2089
-29.95	0.000	4.54	0.99	5.62	Extranjera	Nacionalidad	NA03	2689
-30.09	0.000	4.16	0.88	5.43	Primaria	Nivel de estudios	ED02	2596
-30.64	0.000	10.23	4.68	11.75	Hasta 1 mes	Duración del contrato	DU02	5622
-31.23	0.000	9.91	4.51	11.68	Financiero-Profesion	Sector de actividad	AC08	5588
-34.00	0.000	1.18	0.22	4.79	Construcción	Sector de actividad	AC05	2291
-34.55	0.000	2.91	0.69	6.11	Hasta 6 meses	Duración del contrato	DU03	2921
-36.68	0.000	1.78	0.42	5.98	Industria 2	Sector de actividad	AC03	2862
-37.06	0.000	6.77	2.87	10.87	2-3 años	Tiempo en la empresa	AN02	5199
-38.01	0.000	0.32	0.07	5.26	Industria 1	Sector de actividad	AC02	2518
-38.19	0.000	3.18	0.94	7.62	Otros servicios	Sector de actividad	AC10	3646
-42.52	0.000	1.92	0.60	8.06	Operadores	Ocupación	OC08	3854
-44.07	0.000	11.07	10.46	24.27	Temporal	Tipo de contrato	CO02	11608
-45.19	0.000	7.86	5.35	17.48	Parcial	Tipo de jornada	JO02	8364
-45.90	0.000	3.06	1.26	10.60	Cualificado industri	Ocupación	OC07	5069
-50.70	0.000	4.46	2.57	14.80	Elementales	Ocupación	OC09	7080
-58.89	0.000	3.22	2.12	16.90	Hasta 1 año	Tiempo en la empresa	AN01	8083
-69.10	0.000	5.55	6.00	27.76	Secundaria 1a	Nivel de estudios	ED03	13279
-80.56	0.000	0.27	0.22	20.98	Comercio	Sector de actividad	AC06	10034
-99.99	0.000	0.00	0.00	0.00	Sin datos	Nivel de estudios	ED07	0
-99.99	0.000	0.00	0.00	0.00	0 a 4 años	Edad	ED01	0
-99.99	0.000	0.00	0.00	0.00	10 a 15 años	Edad	ED03	0
-99.99	0.000	0.00	0.00	0.00	Sin datos	Tiempo en la empresa	AN06	0
-99.99	0.000	0.00	0.00	0.00	Sin datos	Tipo de empresa asalariados	EM04	0
-99.99	0.000	0.00	0.00	0.00	No asalariado	Tipo de empresa asalariados	EM03	0
-99.99	0.000	0.00	0.00	0.00	Sin datos	Tipo de jornada	JO03	0
-99.99	0.000	0.00	0.00	0.00	5 a 9 años	Edad	ED02	0
-99.99	0.000	0.00	0.00	0.00	Sin datos	Sector de actividad	AC11	0
-99.99	0.000	0.00	0.00	0.00	Sin datos	Tipo de contrato	CO03	0
-99.99	0.000	0.00	0.00	0.00	65 o más años	Edad	ED14	0
******	0.000	6.61	19.62	76.24	Privada	Tipo de empresa asalariados	EM02	36468

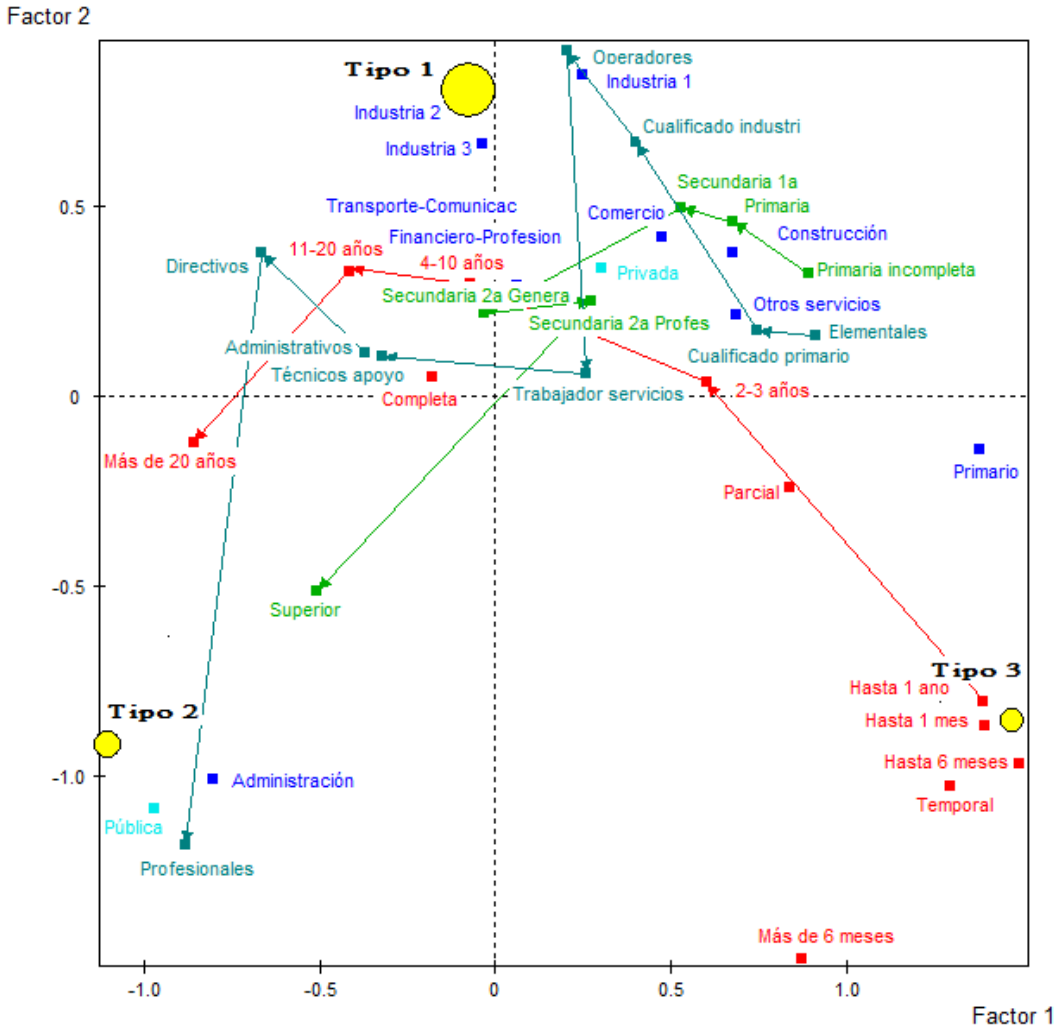
CLUSTER 3 / 3

T.VALUE	PROB.	PERCENTAGES GRP/CAT	PERCENTAGES CAT/GRP	PERCENTAGES GLOBAL	CHARACTERISTIC CATEGORIES	OF VARIABLES	IDEN	WEIGHT
				21.92	**CLUSTER 3 / 3**		**aa3a**	**10485**
199.48	0.000	88.72	98.23	24.27	Temporal	Tipo de contrato	CO02	11608
131.89	0.000	82.05	63.25	16.90	Hasta 1 año	Tiempo en la empresa	AN01	8083
118.74	0.000	89.63	48.06	11.75	Hasta 1 mes	Duración del contrato	DU02	5622
92.43	0.000	97.09	27.05	6.11	Hasta 6 meses	Duración del contrato	DU03	2921
63.06	0.000	77.94	19.34	5.44	Más de 6 meses	Duración del contrato	DU04	2602
50.29	0.000	43.91	35.03	17.48	Parcial	Tipo de jornada	JO02	8364
47.34	0.000	69.63	13.58	4.28	20 a 24 años	Edad	ED05	2045
39.02	0.000	25.77	89.64	76.24	Privada	Tipo de empresa asalariados	EM02	36468
36.51	0.000	39.46	26.65	14.80	Elementales	Ocupación	OC09	7080
36.48	0.000	45.89	18.52	8.85	25 a 29 años	Edad	ED06	4232
31.05	0.000	58.81	8.21	3.06	Primario	Sector de actividad	AC01	1464
25.90	0.000	36.74	18.22	10.87	2-3 años	Tiempo en la empresa	AN02	5199
21.32	0.000	28.55	36.16	27.76	Secundaria 1a	Nivel de estudios	ED03	13279
20.34	0.000	38.75	9.94	5.62	Extranjera	Nacionalidad	NA03	2689
18.09	0.000	38.19	8.35	4.79	Construcción	Sector de actividad	AC05	2291
12.58	0.000	30.53	10.62	7.62	Otros servicios	Sector de actividad	AC10	3646
10.59	0.000	27.90	13.49	10.60	Cualificado industri	Ocupación	OC07	5069
10.13	0.000	25.69	24.59	20.98	Comercio	Sector de actividad	AC06	10034
9.45	0.000	29.66	7.34	5.43	Primaria	Nivel de estudios	ED02	2596
9.22	0.000	25.25	25.34	22.00	Trabajador servicios	Ocupación	OC05	10524
9.03	0.000	26.90	13.47	10.97	30 a 34 años	Edad	ED07	5250
7.84	0.000	26.47	12.06	9.98	Secundaria 2a Profes	Nivel de estudios	ED05	4775
4.07	0.000	27.42	2.57	2.05	Española y doble nac	Nacionalidad	NA02	981
-2.60	0.005	20.10	6.29	6.86	Transporte-Comunicac	Sector de actividad	AC07	3283
-2.98	0.001	19.29	3.84	4.37	Industria 3	Sector de actividad	AC04	2089
-3.96	0.000	19.98	11.77	12.91	Secundaria 2a Genera	Nivel de estudios	ED04	6176
-4.18	0.000	19.76	10.53	11.68	Financiero-Profesion	Sector de actividad	AC08	5588
-6.35	0.000	17.26	4.71	5.98	Industria 2	Sector de actividad	AC03	2862
-10.22	0.000	17.45	12.10	15.21	40 a 44 años	Edad	ED09	7274
-11.85	0.000	15.63	7.57	10.62	Técnicos apoyo	Ocupación	OC03	5081
-15.44	0.000	15.27	10.65	15.29	45 a 49 años	Edad	ED10	7314
-17.14	0.000	13.42	7.15	11.68	Administrativos	Ocupación	OC04	5587
-19.34	0.000	2.78	0.31	2.48	Directivos	Ocupación	OC01	1188
-19.60	0.000	13.21	8.53	14.14	50 a 54 años	Edad	ED11	6766
-19.91	0.000	20.77	87.50	92.33	Española	Nacionalidad	NA01	44166
-21.40	0.000	6.60	1.57	5.23	60 a 64 años	Edad	ED13	2501
-23.47	0.000	9.90	4.72	10.45	55 a 59 años	Edad	ED12	5001
-26.12	0.000	11.92	9.77	17.96	Profesionales	Ocupación	OC02	8590
-28.21	0.000	15.82	30.86	42.76	Superior	Nivel de estudios	ED06	20453
-30.89	0.000	13.19	17.68	29.39	Administración	Sector de actividad	AC09	14061
-37.50	0.000	11.07	14.07	27.86	4-10 años	Tiempo en la empresa	AN03	13326
-39.02	0.000	9.55	10.36	23.76	Pública	Tipo de empresa asalariados	EM01	11368
-50.29	0.000	17.26	64.97	82.52	Completa	Tipo de jornada	JO01	39472
-61.33	0.000	3.27	3.39	22.71	11-20 años	Tiempo en la empresa	AN04	10862
-69.67	0.000	1.09	1.08	21.67	Más de 20 años	Tiempo en la empresa	AN05	10366
-99.99	0.000	0.00	0.00	0.00	10 a 15 años	Edad	ED03	0
-99.99	0.000	0.00	0.00	0.00	Sin datos	Nivel de estudios	ED07	0
-99.99	0.000	0.00	0.00	0.00	Sin datos	Tipo de empresa asalariados	EM04	0
-99.99	0.000	0.00	0.00	0.00	65 o más años	Edad	ED14	0
-99.99	0.000	0.00	0.00	0.00	0 a 4 años	Edad	ED01	0
-99.99	0.000	0.00	0.00	0.00	Sin datos	Sector de actividad	AC11	0
-99.99	0.000	0.00	0.00	0.00	No asalariado	Tipo de empresa asalariados	EM03	0
-99.99	0.000	0.00	0.00	0.00	Sin datos	Tipo de jornada	JO03	0
-99.99	0.000	0.00	0.00	0.00	5 a 9 años	Edad	ED02	0
-99.99	0.000	0.00	0.00	0.00	Sin datos	Tipo de contrato	CO03	0
-99.99	0.000	0.00	0.00	0.00	Sin datos	Tiempo en la empresa	AN06	0
******	0.000	0.51	1.77	75.73	Indefinido	Duración del contrato	DU01	36228
******	0.000	0.51	1.77	75.73	Indefinido	Tipo de contrato	CO01	36228

De la lectura de estas tablas extraemos los siguientes perfiles descriptivos básicos. El tipo 1, con el 52% de los asalariados, reúne a los trabajadores indefinidos de las empresas privadas y públicas, con cierta antigüedad en el mercado de trabajo y de diversos sectores de actividad, sobre todo del comercio, industria y financiero. Se trata de empleados con cualificaciones intermedias o bajas. Se corresponde con un segmento de empleo característico de los que se ha venido en llamar segmento primario dependiente o inferior, estable y de cualificación media. El tipo 2, con el 26% de los casos, recoge fundamentalmente el empleo público de trabajadores altamente cualificados mayoritariamente contratados de forma indefinida, aunque no solamente, y con la mayor antigüedad, donde destaca en particular la mayor presencia femenina. Se trata pues del perfil característico del segmento primario independiente, cualificado y estable. Por último, el tipo 3, con el 22% de los trabajadores, identifica al perfil que se suele denominar como segmento secundario, caracterizado sobre todo por la precariedad del empleo y la baja cualificación en puestos de trabajo del sector privado,

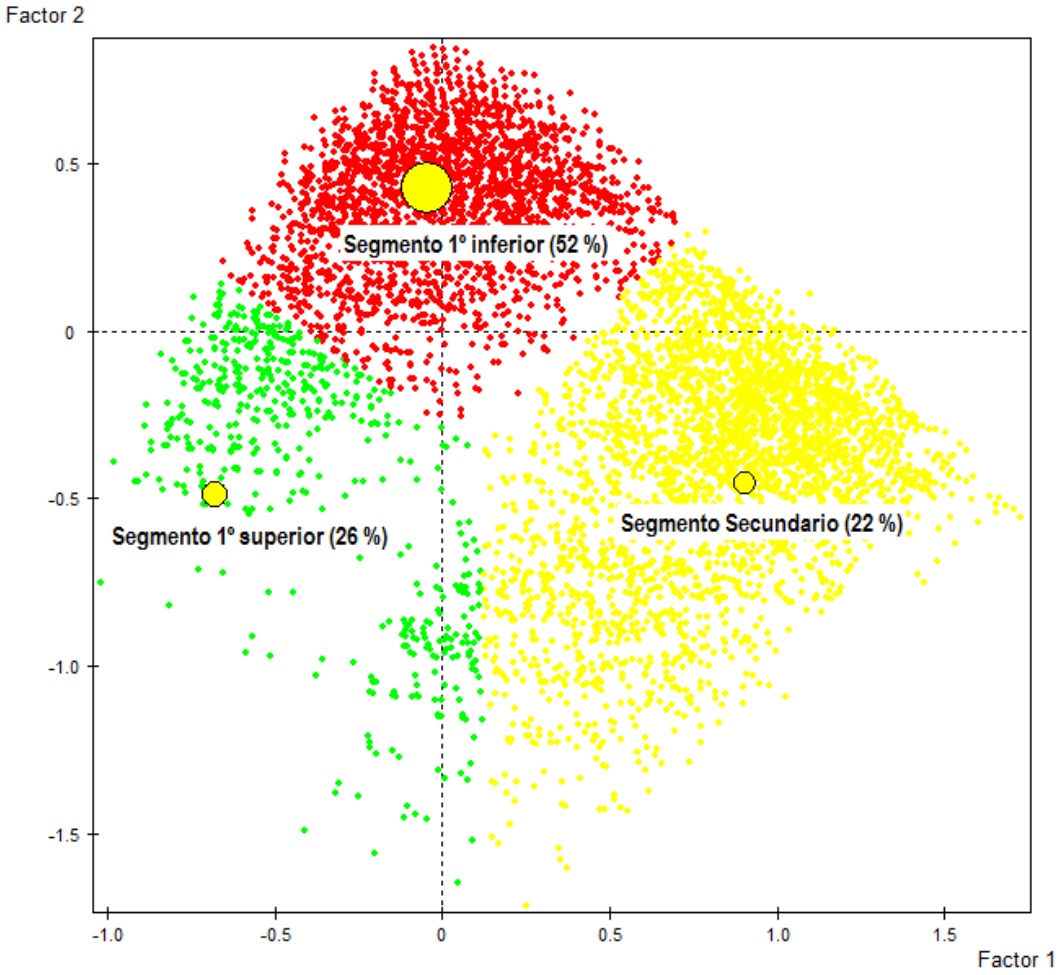
en la construcción, el comercio y otros servicios, que son ocupados en mayor medida por personas jóvenes de ambos sexos y la población inmigrante.

Los gráficos factoriales muestran igualmente los perfiles que acabamos de resumir. Se presentan dos gráficos. En eñ primero representamos los centros de cada uno de los tres tipos obtenidos en el espacio factorial junto a las variables activas.

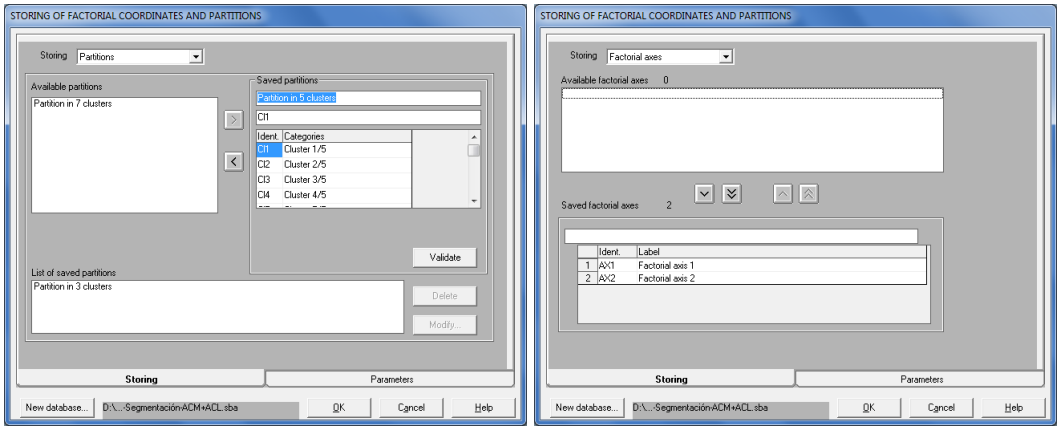


El gráfico siguiente representa a todos los individuos en el espacio factorial de acuerdo con su clasificación en cada uno de los tres tipos. Los tres colores del gráfico sitúan a cada individuo en el tipo al que pertenece: el segmento primario, inferior o superior, o al segmento secundario.

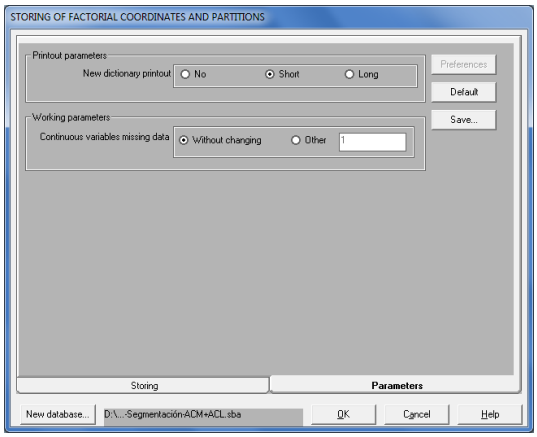
Las otras clasificaciones en 5 y 7 grupos pueden ser estudiadas igualmente como tipologías más detalladas que precisan la visión más sintética en 3 tipos. De ingual forma una clasificación en dos diferenciando el segmento primario del secundario poder argumentarse como forma explicativa de la estructuración del mercado de trabajo.



El programa de instrucciones con el que se ha realizado el análisis de correspondencias en combinación con el análisis de clasificación se completa con un procedimiento adicional que hemos incluido a través de Insert method: el método ESCLA, destinado a guardar las variables factoriales y las tipologías en una base de datos en formato de SPAD que luego también se puede exportar hacia otro software. En el desplegable del cuadro de diálogo inicial (Storing) elegiremos particiones y factores sucesivamente y las variables que aparecen las pasaremos a la derecha y las etiquetaremos y validaremos.



En New Database detallaremos el nombre y la localización del fichero que se guardará. Finalmente en Parameters se concretan dos opciones en relación a la presentación del diccionario de los datos y los valores perdidos de las variables cuantitativas.



A continuación se propone la realización de un ejercicio de análisis tipológico con los datos de la matriz Conciliación.sba donde tras realizar un análisis de correspondencias múltiples se aplica un análisis de clasificación jerárquico. Los datos se emplearon en una investigación publicada por López-Roldán y Lozares (2007) donde se examina, en primer lugar, el grado de asociación o independencia que guardan entre sí el ámbito socioproductivo y las condiciones sociofamiliares y, en segundo lugar, la incidencia que la conjugación de dichos ámbitos tiene sobre las prácticas y responsabilidades del hogar, así como sobre la dificultad para llevarlas a cabo. Los resultados son reveladores de las tensiones y contradicciones que provienen de conjugar, en la vida de los trabajadores y las trabajadoras, individualmente y como pareja, las exigencias que provienen de dichos ámbitos y, por tanto, de conciliar sus estrategias, proyectos, actividades y la posibilidad de acuerdos.

El análisis se basa en los datos de la encuesta llevada a cabo en una empresa multinacional de producción de medios de comunicación audiovisual a partir de un cuestionario administrado a 262 trabajadores y trabajadoras de las secciones de producción, inserción y diseño, sobre una plantilla de unos 1.200 trabajadores.

Las variables utilizadas en el análisis se adjuntan a continuación. Corresponden a las variables V001 a V023 que se considerarán como variables activas todas ellas. Para analizar los datos se precisa ponderarlos por el diseño de la muestra y para respetar la proporción de cada sección de la empresa. En la pestaña Weighting clicaremos sobre la opción de Weighting variable y elegiremos la variable V024. El resto de las opciones las mantendremos por defecto.

Realizaremos un análisis de clasificación jerárquico ascendente con el método Ward (opción `RECIP`) con la optimización o consolidación a partir de centros móviles que se aplica por defecto. Como resultado del análisis se consideraron 6 tipos de trabajadores de la empresa desde el punto de vista de la conciliación. Los resultados del análisis se pueden consultar en López-Roldán y Lozares (2007: 131-136).

Variables utilizadas en el análisis de conciliación

1) Ámbito del trabajo productivo en la empresa

a) Variables temporales de trabajo productivo cotidiano: sistema de horario cotidiano en la empresa vinculado a la edad y al tiempo en la empresa.
Horario de trabajo: Partido / Rotatorio / Mañana / Tarde
Antigüedad en la empresa: 1974-92 / 1993-97 / 1998-2000
Edad del trabajador: 18-24 / 25-29 / 30-45

b) Variables adicionales de caracterización del puesto de trabajo
Sección de la empresa: Diseño / Inserción / Producción
Categoría profesional: Técnico superior o medio / Operario, Administrativo
Nivel de Estudios: Primarios-EGB / BUP / FP1 / FP2 / Universitarios
Tipo de contrato: Indefinido / Temporal
Género: Varón / Mujer

2) Condiciones sociofamiliares

a) Variables que definen el tiempo de ciclo de vida expresión del tiempo social
Tipo de convivencia: Vive con padres / Solo / En pareja / Pareja con hijos / Otras
Presencia de personas dependientes: Ninguna / Personas mayores / Hijos

b) Variables temporales familiares vinculadas al trabajo productivo
Personas Ocupadas: Ocupado él-ella / Pareja ocupada / Él-ella+Padres / Otras
Coincidencia de horarios pareja: Coinciden / Parcialmente / No coinciden / No pareja

3) Ámbito del trabajo reproductivo

a) Percepción de las dificultades del trabajo reproductivo
Tareas del hogar: Dificultades para realizarlas / No dificultades
Cuidado de los hijos: Dificultades para realizarlas / No dificultades

b) Reparto del trabajo reproductivo
Quién cocina: Él-ella / Ambos / Él-ella+Pareja / Él-ella+Padres / Otras situaciones
Quién limpia: Él-ella / Ambos / Él-ella+Pareja / Él-ella+Padres / Otras situaciones
Quién compra: Él-ella / Ambos / Él-ella+Pareja / Él-ella+Padres / Otras situaciones
Quién bricolaje: Él-ella / Ambos / Él-ella+Pareja / Él-ella+Padres / Otras situaciones
Quién cuida hijos: Él-ella / Él-ella+Pareja / Él-ella+Padres / Otras / No hijos

4) Ámbito del tiempo libre (no trabajo): posibilidad disfrutar del tiempo de ocio

Participación en asociaciones: Dificultades para participar / No dificultades
Ocio familiar: Dificultades para disfrutarlo / No dificultades
Ocio individual: Dificultades para disfrutarlo / No dificultades
Ocio amistades: Dificultades para disfrutarlo / No dificultades

Se adjuntan a continuación tres gráficos factoriales: de las categorías de las variables consideradas en el espacio de los dos primeros factores, el de los centros de cada tipo o *cluster* obtenido y de éstos en el centro de los individuos.

Gráfico de las categorías en el espacio de los factores 1 y 2

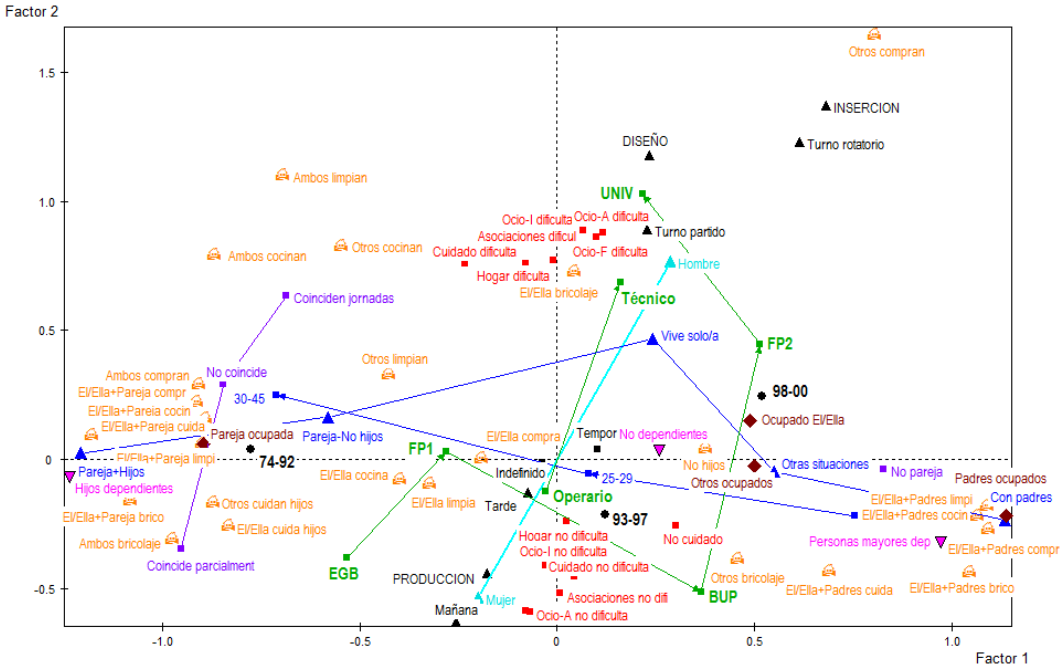


Gráfico de las categorías y centros de los grupos en el espacio de los factores 1 y 2

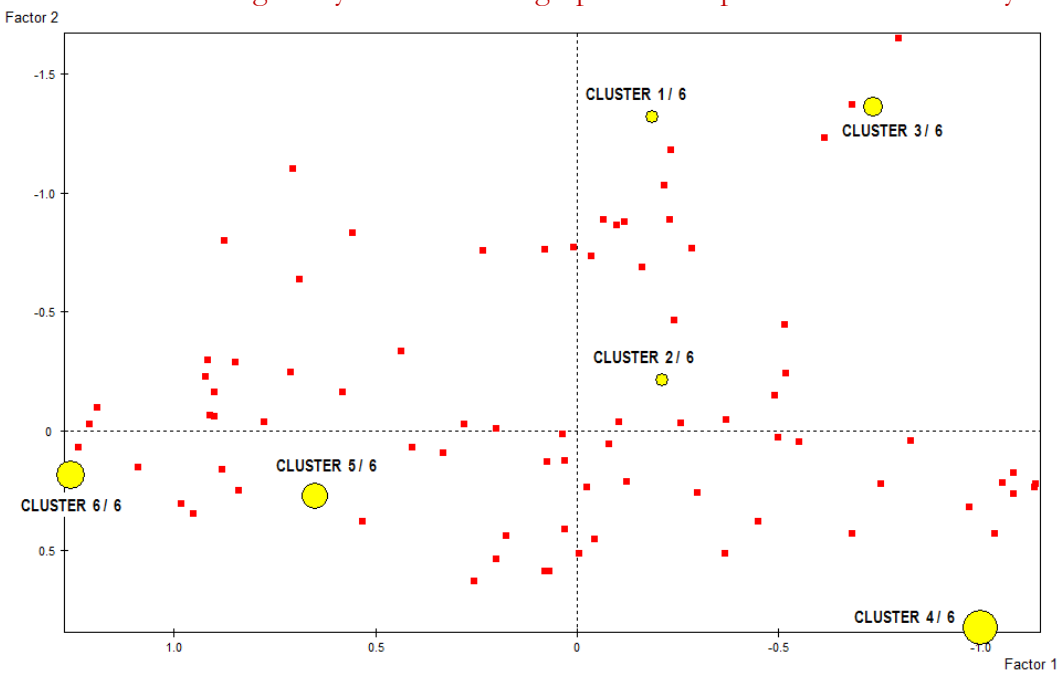
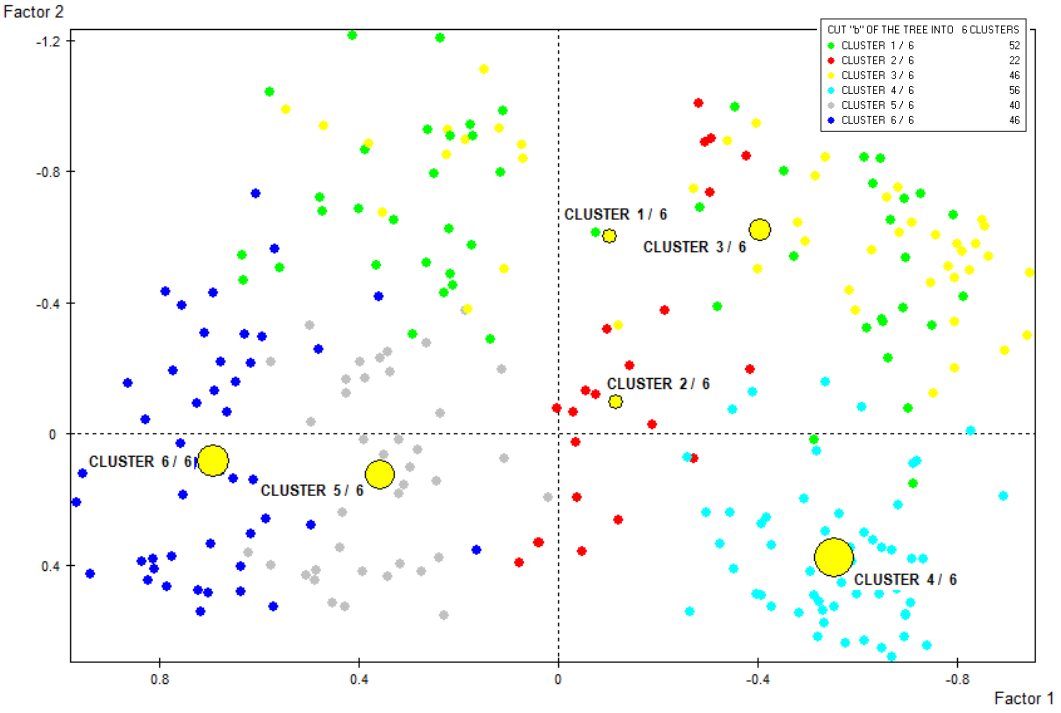
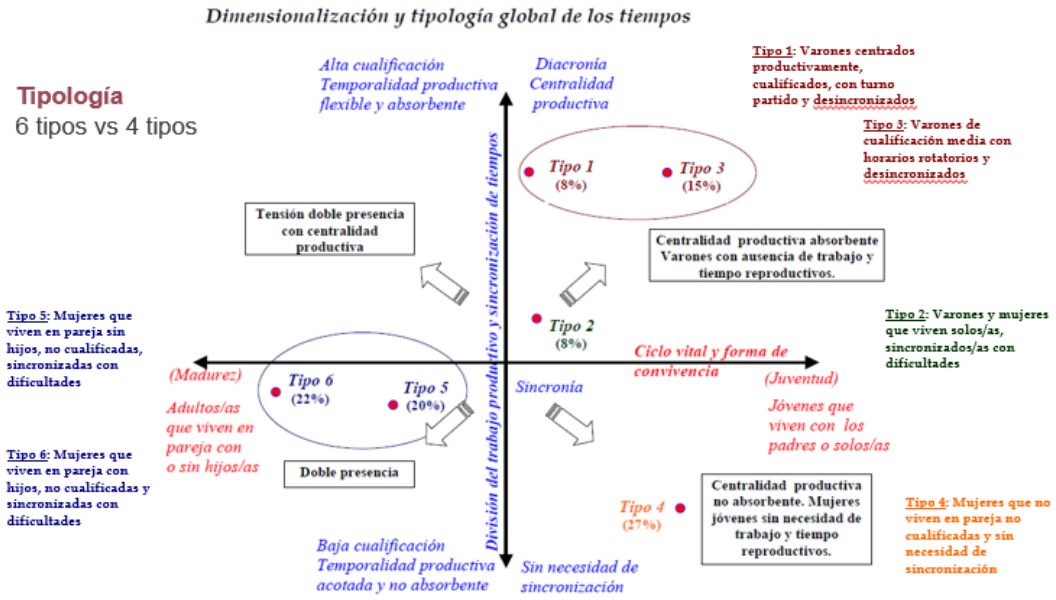


Gráfico de los individuos y centros de los grupos en el espacio de los factores 1 y 2



De forma alternativa los resultados del análisis en el espacio de los dos principales factores se pueden representar así:



6. Análisis de clasificación con R

Por desarrollar

7. Bibliografía

Aldenderfer, M. S..; Blashfield, R. K. (1987). *Cluster Analysis*. Beverly Hills: Sage Publications.

Anderberg, M. R. (1973). Cluster Analysis for Applications. New York: Academic Press.

Arabie, P.: Hubert, L. J.; De Soete, G. (Eds.) (1996). Clustering and classification. Singapore: World Scientific.

Bailey, K. D. (1975). Cluster analysis. En D. R. Heise (Ed.), *Sociological Methodology*. San Francisco, California: Jossey-Bass.

Bailey, K. D. (1983). Sociological Classification and cluster analysis. *Quality and Quantity*, 17, 251-268.

Bailey, K. D. (1989). Constructing Typologies trough Cluster Analysis. *Bulletin de Méthodologie Sociologique*, 25, 17-28.

Bailey, K. D. (1994). Typo*logies and Taxonomies. An Introduction to Classification Techniques*. Thousand Oaks (California): Sage.

Bailey, K. D. (2004). Typology construction: Methods and issues. En: *Encyclopedia of Social Measurement*, editada por K. Kempf-Leonard. Vol. 1-3. San Diego, CA: Academic Press.

Barbut, M.; Monjardet, B. (1970). *Ordre et classification. Algèbre et combinatoire*. Hachette, Paris.

Barton, A. H. (1985). Concepto de espacio de atributos en Sociología. En R. Boudon y P. F. Lazarsfeld, *Metodología de las Ciencias Sociales. I. Conceptos e índices*. Barcelona: Laia, 195-219.

Bécue, M.; Valls, J. *Manual de introducción a los métodos factoriales y clasificación con SPAD*. Bellaterra: Servei d'Estadística de la UAB. http://sct.uab.cat/estadistica/sites/sct.uab.cat.estadistica/files/manualSPAD.pdf

Benzécri, J. P. (1973). L'Analyse des données. I. *La taxonomie*. Paris: Dunod.

Bertier, P.; Bourouche, J.-M. (1983). *Analyse des données multidimensionnelles*. Paris: PUF.

Calinski, T.; Harabasz J. (1974) A dendrite method for cluster analysis. *Commun Stat Theory Methods*, 3, 1, 1–27

Capecchi, V. (1964). Une méthode de classification fondée sur l'entropie. *Revue Française de Sociologie*, V, 3, 290–306.

Capecchi, V. (1966). Typologies in Relation to Mathematical Models. *Ikon,* Suplemento, 58, julio-septiembre, 1–62.

Capecchi, V. (1968). On the Definition of Typology and Classification in Sociology. *Quality and Quantity*, 2, 1-2, enero, 9-30.

Cea d'Ancona, M. A. (2002). Análisis de conglomerados. En M. A. Cea d'Ancona, *Análisis multivariable. Teoría y práctica en la investigación social*. Madrid: Síntesis, 230-321.

Chandon, J. L.; Pinson, S. (1981). *Analyse typologique. Théories et applications*. Paris: Masson.

Chiu, T.; Fang, D.; Chen, J.; Wang, Y.; Jeris C (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. *Proceedings of the 7th ACM SIGKDD international conference in knowledge discovery and data mining*, Association for Computing Machinery, San Francisco, CA, 263–268.

CISIA-CERESTA (2001). *Introduction à SPAD Version 5.0. Manuel de Prise en Main*. Montreuil: CISIA-CERESTA

CISIA-CERESTA (2001). *Système SPAD pour Windows Version 5.0. SPAD-Base. Aide à l'interprétation*. Montreuil: CISIA-CERESTA.

Cohen, N.; Gómez, G. (2011). Las tipologías y sus aportes a las teorías y la producción de datos. *Revista Latinoamericana de Metodología de la Investigación Social*, 1, abril-septiembre, 36-46. http://www.relmis.com.ar/ojs/index.php/relmis/article/view/9/12

COHERIS-SPAD (2007). *SPAD7.0. Introduction à SPAD*. Guide de l'utilisateur. Courbevoie: SPAD. http://tic-recherche.crifpe.ca/docs/guides/fr/SPAD7_guide.pdf

Cuadras, C. M. (1989). Distancias estadísticas. *Estadística Española*, 30, 119, setiembre-diciembre, 295-378.

Cuadras, C. M. (2012). *Nuevos métodos de análisis multivariante*. Barcelona: CMC Editions. http://www.ub.edu/stat/personal/cuadras/metodos.pdf

De Martinelli, G. (2011). De los conceptos a la construcción de los tipos sociales agrarios. Una mirada sobre distintos modelos y las estrategias metodológicas. *Revista Latinoamericana de Metodología de la Investigación Social*, 2, octubre-marzo, 24-43. http://www.relmis.com.ar/ojs/index.php/relmis/article/view/22/19

Dice, L. R. (1945). Measure of the amount of ecologic associations between species. *Ecology*, 26, 277-302.

Diday, E. (1971). La méthode des Nuées Dynamiques. *Revue de Statistique Appliquée*, 19, 2, 19-34.

Dimitriadou, E., Dolnicar, S.; Weingessel, A. (2002) An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, 67, 137–159.

Domínguez, M.; López-Roldán, P. (1996). La construcció de tipologies: procés i tècniques d'anàlisi de dades. *Papers. Revista de Sociologia*, 48, 31-39. http://ddd.uab.cat/pub/papers/02102862n48p31.pdf

Domínguez, M.; Sánchez Miret, C. (1996). Aspectes metodològics i tècnica d'anàlisi de les dades per a l'estudi dels grups i les classes socials a la Regió Metropolitana de Barcelona. *Papers. Revista de Sociologia*, 48, 59-69. http://papers.uab.cat/article/view/v48-dominguez-sanchez/pdf-ca

Driver, H. E.; Kroeber, A. L. (1932). Quantitative expression of cultural relationships. *University of California Publications in American Archeology and Ethnology*,31, 211-256.

Edwards, W.F.; Cavalli-Sforza, L. L. (1965). A method for cluster analysis. *Biometrics*, 21, 362-375.

Elías Pardo, C.; Del Campo, P. C. (2007). Combinación de métodos factoriales y de análisis de conglomerados en R: el paquete factoclass. *Revista Colombiana de Estadística*, 30, 2, 231-245. http://www.kurims.kyoto-u.ac.jp/EMIS/journals/RCE/V30/v30-2body/v30n2a06PardoDelCampo.pdf

Etxeberría, J.; García, E.; Gil, J.; Rodríguez, G. (1995). Análisis de datos y textos. Madrid: RA-MA.

Everitt, B. S. (1983). Cluster Analysis. En D. McKay, N. Schofield i P. Whiteley, *Data Analysis and Social Sciences*. London: Frances Pinter Publishers, 226-255.

Everitt, B. S. et al. (2011). *Cluster Analysis*. Chichester, UK: John Wiley & Sons.

Fachelli, S. (2009). *Nuevo modelo de estratificación social y nuevo instrumento para su medición. El caso argentino*. Tesis Doctoral. Bellaterra: Universitat Autònoma de Barcelona. http://tdx.cat/handle/10803/5149

Fachelli, S. (2010). Trayectorias de los hogares argentinos según estrato social entre 1997 y 2006. *Revista Latinoamericana de Estudios del Trabajo*, 23-24, 81-112. http://relet.iesp.uerj.br/Relet_23-24/art5.pdf

Fachelli, S. (2013). ¿La crisis aumenta las diferencias entre estratos sociales? La medición del cambio social en Argentina. *EMPIRIA. Revista de Metodología de Ciencias Sociales*, 22, enero-junio, 13-46. http://dialnet.unirioja.es/descarga/articulo/4118170.pdf

Fachelli, S.; Goicoechea, M. E.; López-Roldán, P. (2015). Trazando el mapa social de Buenos Aires: dos décadas de cambios en la Ciudad. *Población de Buenos Aires*, 12, 21, 7-39. http://ddd.uab.cat/pub/artpub/2015/132095/pobbueair_a2015n21p7iSPA_postprint.pdf

Fachelli, S.; López, N.; López-Roldán, P.; Sourrouille, F. (2012). *Desigualdad y diversidad en América Latina: hacia un análisis tipológico comparado*. Buenos Aires: SITEAL, Instituto Internacional de Planeamiento de la Educación (UNESCO-OEI). Libros digitales, 2. http://www.siteal.org/sites/default/files/siteal_libro_digital_desigualdad_y_diversidad.pdf http://pagines.uab.cat/plopez/sites/pagines.uab.cat.plopez/files/SITEAL-UBA.pdf

Fachelli, S.; López-Roldán, P. (2010). An attempt to measure social stratification and social changes in terms of distances. *XVII ISA World Congress of Sociology*, 11-17 de Julio, Göteborg (Suecia). https://ddd.uab.cat/record/113791 http://pagines.uab.cat/plopez/sites/pagines.uab.cat.plopez/files/Estratos-UBA.pdf

Faggiano, M. P. (2012). *Gli usi della tipologia nella ricerca sociale empírica*. Milano: Franco Angeli Edizioni.

Fernández, O. (1991). El análisis de cluster: aplicación, interpretación y validación, *Papers. Revista de Sociologia*, 37, 65-76. http://ddd.uab.cat/pub/papers/02102862n37p65.pdf

Fichet, B. et al. (2011). *Classification and Multivariate Analysis for Complex Data Structures*. Berlin: Springer-Verlag.

Goicoechea, M. E. (2014). *El mapa social de Buenos Aires (2001)*. http://ddd.uab.cat/pub/trerecpro/2014/117077/TFG_megoicoetxea.pdf

Goldemberg, J., Fisherman, J. y Torres, H. A. (1967). Déficit habitacional y tendencias ecológicas en la Ciudad de Buenos Aires. *Revista SUMMA*, 9, agosto.

Gower, J. C. (1967). A comparison of some methods of cluster analysis. *Biometrics*, 23, 623-637.

Hamann, U. (1961). Merkmalbestand und Verwandtschafts-beziehungen der Farinosae. Ein Beitragzum System der Monokotyledonen. *Willdenowia*, 2, 639-768.

Hartigan, J. A. (1975). *Clustering algorithms*. New York: John Wiley.

Hair, J. F. et al. (2011). *Multivariate Data Analysis*. Upper Saddle River: Prentice Hall.

Hernández, L. (2001). *Técnicas de taxonomía numérica*. Madrid: La Muralla.

Hempel, C. G. (1952). Typological Methods in the Natural and the Social Sciences. *Proceedings of the American Philosophical Association: Eastern Division*, 1, 656-686.

Hempel, C. G. (1965). *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*. Free Press, New York.

Hempel, C. G. (1979). Métodos tipológicos en las ciencias naturales y sociales. En C.G. Hempel, *La explicación científica. Estudios sobre filosofía de la ciencia*. Paidós, Buenos Aires, 159-175.

Herrera-Usagre, M. (2011). El consumo cultural en España. Una aproximación al análisis de la estratificación social de los consumos culturales y sus dificultades metodológicas. *EMPIRIA. Revista de Metodología de Ciencias Sociales*, 22, julio-diciembre, 141-172. http://e-spacio.uned.es/fez/eserv.php?pid=bibliuned:Empiria-2011-22-5060&dsID=Documento.pdf

Husson, F.; Lê, S.; Pagès, J. (2011). *Exploratory Multivariate Analysis by Example using* R. London: Chapman & Hall. http://factominer.free.fr/book

Itzcovich, G.; Sourrouille, F. (2012). *Condiciones sociales, configuraciones familiares y vínculos de escolarización en adolescentes de 15 a 17 años. Aproximación desde una perspectiva relacional*. SITEAL, Instituto Internacional de Planeamiento de la Educación (UNESCO-OEI). Cuaderno 13. http://www.siteal.org/sites/default/files/cuaderno13_20121002.pdf

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 44, 223-270.

Jambu, M.; Lebeaux, M. O. (1978). *Classification automatique pour l'analyse des données*. Paris: Dunod.

Kulczynski, S. 1927. Die Pflanzenassoziationen der Pienienen. *Bull. Intern. Acad. Pol. Sci. Lett. Cl. Sci. Math. Nat., B (Sci. Nat.)*, Suppl. 2, 57-203.

Lazarsfeld, P. F. (1937). Some remarks on the typological procedure in social research, *Zeitschrift für Sozialforschung*, 6, 119-139.

Lazarsfeld, P. F. (1985). De los conceptos a los índices empíricos. En: *Metodología de las Ciencies Sociales. I. Conceptos e Indices*, editado por R. Boudon y P.F. Lazarsfeld, Barcelona: Laia, 35-62.

Lazarsfeld, P. F.; Barton, A. H. (1951). Qualitative Measurement in the Social Sciences: Classification, Typologies and Indices. En D. Lerner y H.D. Lasswell, *The Policy Sciences*. Stanford: Stanford University Press, 155-192.

Lebart, L.; Morineau, A.; Fenelon, J. P. (1985). *Tratamiento estadístico de datos. Métodos y programas*. Barcelona: Marcombo.

Lebart, L.; Morineau, A.; Piron, M. (1997). *Statistique exploratoire multidimensionnelle*. Paris: Dunod.

López-Roldán, P. (1994). La construcción de tipologías en Sociología: propuesta metodológica de construcción, análisis y validación. Aplicación al estudio de la segmentación del mercado de trabajo en la Regió Metropolitana de Barcelona, Tesi Doctoral, Universitat Autònoma de Barcelona, Bellaterra (Barcelona). http://www.tdx.cat/handle/10803/5135

López-Roldán, P. (1996a). La construcción de tipologías: metodología de análisis. *Papers. Revsita de Sociologia*, 48, 9-29. http://ddd.uab.cat/pub/papers/02102862n48p9.pdf

López-Roldán, P. (1996b). La construcción de una tipología de segmentación del mercado de trabajo. *Papers. Revista de Sociologia*, 48, 41-58. http://ddd.uab.cat/pub/papers/02102862n48p41.pdf

López-Roldán, P. (2011). La Muestra Continua de Vidas Laborales: posibilidades y limitaciones. Aplicación al estudio de la ocupación de la población inmigrante. *Metodología de Encuestas*, 13, 7-32. http://casus.usal.es/pkp/index.php/MdE/article/view/1010

López-Roldán, P. (2012). El proceso metodológico de construcción tipológica de las clases sociales. En: *Barcelona: de la necessidad a la libertad. Las clases sociales en los albores del siglo XXI*, editado por M. Subirats. Barcelona: UOC Ediciones, 423-431.

López-Roldán, P.; Fachelli, S. (2015). *Metodología de construcción de tipologías para el análisis de la realidad social*. Bellaterra: Universitat Autònoma de Barcelona. https://ddd.uab.cat/record/118082

López-Roldán, P.; Lozares, C. (2008). La construcción de la muestra. En Institut d'Estudis Regionals i Metropolitans de Barcelona, *El trabajo de campo de la Encuesta de condiciones de vida y hábitos de la población de Cataluña, 2006*. Barcelona: IERMB, 17-39. http://www.iermb.uab.es/htm/publicacions.asp?idPubCat=13&idPub=115

López-Roldán, P.; Lozares, C. (2007). La conciliación entre las exigencias del ámbito productivo y las condiciones socio-familiares: estudio de caso de una empresa. *Papers. Revista de Sociologia*, 83, 123-144. http://ddd.uab.cat/pub/papers/02102862n83p123.pdf

López-Roldán, P.; Lozares, C.; Domínguez, M. (2000). Disseny i construcció d'una mostra estratificada a partir de dades censals. *Qüestiió*, 24, 1, 111-136. http://www.raco.cat/index.php/Questiio/article/viewFile/143995/195695

López-Roldán, P.; Miguélez, F.; Lope, A. (1998). La segmentación laboral: hacia una tipología del ámbito productivo. *Papers. Revista de Sociologia*, 55, 45-77. http://ddd.uab.cat/pub/papers/02102862n55p45.pdf

Lorr, M. A. (1968). A review and classification of typological procedures. *Paper read at the meeting of the American Psychological Association*, San Francisco, California.

Lozares, C. (1990). La tipología en Sociología: más allá de la taxonomía. *Papers. Revista de Sociologia*, 34, 139-164. http://ddd.uab.cat/pub/papers/02102862n34/02102862n34p139.pdf

Lozares, C.; Domínguez, M. (1993). Tratamiento multivariado de subpoblaciones en una gran encuesta social: la construcción de zonas sociales. *Papers. Revista de Sociologia*, 48. P. 71-87. http://papers.uab.cat/article/view/v48-lozares-dominguez/pdf-es

Lozares, C.; López-Roldán, P.; (1991). El muestreo estratificado por análisis multivariado. En M. Latiesa, *El pluralismo metodológico en la investigación social: ensayos típicos*. Granada: Universidad de Granada, 107-160.

Lozares, C.; López-Roldán, P. (2000). *Anàlisi multivariable de dades estadístiques*. Bellaterra (Barcelona): Universitat Autònoma de Barcelona.

Lozares, C.; López-Roldán, P.; Borràs, V. (1998). La complemetariedad del log-lineal y el análisis de correspondencias en la elaboración y análisis de tipologías. *Papers. Revista de Sociologia*, 55, 79-93. http://ddd.uab.cat/pub/papers/02102862n55p79.pdf

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Science of India*, 12, 49-55.

Marradi, A. (1990). Classification, typology, taxonomy. *Quality & Quantity*, 24, 129-157.

Marradi, A. (2009). Clasificación. En: *Diccionario Crítico de Ciencias Sociales Terminología Científico-Social*, dirigido por Román Reyes. Madrid-México: Ed. Plaza y Valdés. http://pendientedemigracion.ucm.es/info/eurotheo/diccionario/C/clasificacion.pdf

Marradi, A. (2007). La tipologia. Desde Aristóteles a las ciencias sociales modernas. *Revista de Ciencia Política y de Relaciones Internacionales*, I, 1, marzo, 4-24.

http://www.palermo.edu/cienciassociales/PDFCienciaPoliticaN1/latipologiapolitica1.pdf

Martínez, E. (1984). Aspectos teóricos del Análisis de Clúster y Aplicación a la caracterización del electorado potencial de un partdo. En J. J. Sánchez Carrión, *Introducción a las técnicas de análisis multivariable aplicadas a las ciencias sociales*. Madrid: Centro de Investigaciones Sociológicas, 165-208.

Mckinney, J. C. (1968). *Tipología constructiva y teoría social*. Buenos Aires: Amorrortu.

McQuitty, L. L. (1961). Typal analysis. *Educational and Psychological Measurement*, 21, 677–697.

McQuitty, L. L. (1966). Similarity analysis by reciprocal pairs for discrete and continuous data. *Educational and Psychological Measurement*, 26, 825-831.

Miguélez, F; López-Roldán, P. (Coord.) (2014). *Crisis, empleo e inmigración en España. Un análisis de las trayectorias laborales*. Bellaterra (Barcelona): Universitat Autònoma de Barcelona.

Miguélez, F.; Martín, A.; de Alós-Moner, R.; Esteban, F.; López-Roldán, P.; Molina, Ó.; Moreno, S. (2012). *Trayectorias laborales de los inmigrantes en España*. Barcelona: Obra Social "la Caixa". http://multimedia.lacaixa.es/lacaixa/ondemand/obrasocial/pdf/Trayectorias_laborales_de_los_inmigrantes_en_Espana.pdf

Milligan, G. W. (1996) Clustering validation: results and implications for applied analyses. En P. Arabie, L. J. Hubert and G. De Soete, *Clustering and Classification*. Singapore: World Scientific, 341–375.

Milligan, G. W.; Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50, 159–179.

Mooi, E.; Sarstedt, M. (2011). Cluster Analysis. En E. Mooi y M. Sarstedt, A Concise Guide to Market Research. Berlin: Springer-Verlag. DOI 10.1007/978-3-642-12541-6_9.

Morineau, A. (1984). Note sur la caractérisation statistique d'une classe et les valeurs-test. *Bulletin Technique du Centre de Statistique et d'Informatique Appliquées*, 2, 1-2, 20-27. http://www.deenov.com/Data/Sites/1/docs/Valeur-Test-critere-de-caracterisation-statistique.pdf

Norusis, M. J. (2011). Cluster Analysis. En: *IBM SPSS Statistics 19 Statistical Procedures Companion*, editado por M. J. Norusis. Addison Wesley, 375-404. http://www.norusis.com/pdf/SPC_v19.pdf

Núñez-Colín, C. A.; Escobedo-López, D. (2011). Uso correcto del análisis clúster en la caracterización de germoplasma vegetal. *Agronomía Mesoamericana*, 22, 2, 415-427. http://www.mag.go.cr/rev_meso/v22n2_415.pdf

Ochiai, A. 1957. Zoogeographic studies on the soleoidfishes found in Japan and its neighbouring regions. *Bull. Jap. Soc. Sci. Fish*, 22, 526-530.

Picón, E.; Varela, J.; Real, E. (2003). Clasificación y segmentación post hoc mediante análisis de conglomerados. En J.-P. Lévy y J. Varela, *Análisis Multivariable para las Ciencias Sociales*. Madrid: Pearson Prentice Hall, 417-450.

Rogers, DG; Tanimoto, T. T. (1960). A computer program for classifying plants. *Science*, 132, 1115-1118.

Rosemburg, H. Ch. (1984). *Cluster Analysis for Researchers*. Belmont (California). Lifetime Learning Publications.

Russel, P. F.; Rao, T. R. (1940). On habitat and association of species of Anopheline larvae in south-eastern Madras. *J. Malar. Inst. India*, 3, 153-178.

Sánchez, C.; (1994). *La definició dels grups socials a la Regió Metropolitana de Barcelona. Un problema teòric i metodològic*. Tesis Doctoral, Universitat Autònoma de Barcelona, Bellaterra (Barcelona).

Sánchez, C.; Domínguez, M. (2001). Anàlisi de l'estructura social de les comarques catalanes a partir de dades censals. Metodologia i primera aproximació als resultats. *Revista Catalana de Sociologia*, 14, 193-213. http://publicacions.iec.cat/repository/pdf/00000024/00000067.pdf

Sneath, P. H. A.; Sokal, R. R. (1973). *Numerical taxonomy. The Principles and Practice of Numerical Classification*. London: Freeman.

Sokal, R. R.; Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38, 1409-1438.

Sokal, R. R.; Sneath, P. H. A. (1963). *Principles of numerical taxonomy*. San Francisco, California: Freeman.

Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske Skrifter*, 5, 1-34.

Subirats, M.; López-Roldán, P.; Sánchez, C. (2010). Clases y grupos sociales en la Región Metropolitana de Barcelona. *Papers. Regió Metropolitana de Barcelona*, 52, 105-120. http://www.iermb.uab.es/htm/descargaBinaria.asp?idRevArt=276

Subirats, M.; López-Roldán, P.; Sánchez, C. (2010). Classes i grups socials a la Regió Metropolitana de Barcelona. *Papers. Regió Metropolitana de Barcelona*, 52, 8-37. http://www.iermb.uab.cat/htm/descargaBinaria.asp?idRevArt=270

Tiryakian, E. A. (1968) Typologies. En *International Encyclopedia of the Social Sciences*, XVI, London: Macmillan, 177-85.

Torres, H. A. (1983). Encuesta sobre la situación habitacional en la Ciudad de Buenos Aires. *Revista Ideas*, UB, 69-89; *Boletín SEDUV*, nº 19.

Torres, H. A. (1993). El mapa social de Buenos Aires (1940-1990). *Serie Difusión*, n° 3, Buenos Aires: SICyT, Facultad de Arquitectura, Diseño y Urbanismo, UBA.

Torres, H. A. (1999). Diagnóstico socioterritorial de la Ciudad de Buenos Aires y su contexto metropolitano. *Serie Documentos de Trabajo*, n° 1, Plan Urbano Ambiental, GCBA.

Tryon, R. C. (1939). *Cluster analysis: Correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality*. Ann Arbor, Michigan: Edwards Brothers.

Tyron, R. C.; Bailey, D. E. (1970). *Clustering analysis*. New York: McGraw-Hill.

Van Cutsem, B. (1994). *Classification and dissimilarity analysis*. New York: Springer.

Volle, M. (1978). *Analyse des données*. Paris: Economica.

Ward, J. H. Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236-244.

Winch, R. F. (1947). Heuristic and Empirical Typologies: A Job for Factor Analysis. *American Sociological Review*, 12, 1, febrero, 68-75.

Wong, M. A. (1982). A hybrid clustering method for identifying high density clusters. *Journal of American Statistical Association*, 77, 380, 841-847.

Yule, G. U. (1911). *An introduction of the theory of statistics*. London: Charles Griffin & Company.

Zubin, J. A. (1938). A technique for measuring likemindedness. *Journal of Abnormal & Social Psychology*, 33, 508-516.