Metodología de Encuestas
Vol 2, Núm 2, 2000 239-260

CREACIÓN DE OBJETOS SIMBÓLICOS A PARTIR DE ENCUESTAS ALMACENADAS EN BASES DE DATOS RELACIONALES

Patricia Calvo, Marina Ayestarán, Cristina Prado, Yolanda Pérez
EUSTAT, Instituto Vasco de Estadística

RESUMEN

Este trabajo tiene como finalidad la aplicación práctica de las nuevas unidades estadísticas, denominadas Objetos Simbólicos, utilizadas últimamente en el campo de la Estadística Oficial. Estas unidades pretenden resumir grandes cantidades de información almacenada en Bases de Datos Relacionales y describir tanto individuos como grupos de población, que sirvan de referencia para un análisis estadístico posterior. Este proceso se ha hecho posible gracias a la utilización de un software específico (SODAS).

El artículo consta de una presentación metodológica, en la que introducimos el concepto y la creación de Objetos y Tablas Simbólicas. Además, se presentan ejemplos con datos reales como son: el registro de Accidentes de Tráfico de 1999 y la encuesta de Servicios a Empresas de 1998. Finalmente, se plantea una aplicación de estos Objetos Simbólicos para la Fusión de Encuestas.

Palabras clave: objetos simbólicos, SODAS, bases de datos relacionales, fusión de encuestas.

Definición y extensiones de los objetos simbólicos

Qué es un objeto simbólico

Tradicionalmente, se obtienen datos de individuos simples y las variables tienen un solo valor o categoría en cada individuo. Sin embargo, a veces, el mundo real es demasiado complejo para ser descrito por estos modelos relativamente simples. Por ello, se introducen los objetos simbólicos que van a resolver problemas de respuesta no unitaria, es decir, los valores que toman las variables pueden ser *no atómicos* (un grupo de valores, un intervalo de valores o una distribución de probabilidad).

Así pues, *un Objeto Simbólico es un modo de representación de datos complejos que surge al analizar grandes ficheros de datos*, frecuentemente utilizados en los Institutos de Estadística.

Necesidad de Objetos Simbólicos

A continuación se describen varios ejemplos que ilustran la necesidad de utilizar objetos simbólicos:

–Para individuos, la variable Y = "Minutos dedicados a la práctica de deporte al día" es una variable que permite una respuesta no unitaria, ya que varía de día a día. Para un individuo k, esa variable puede expresarse de una forma no clásica:

Y(k) = [20,60] o

Y(k) = {20 minutos (0.15), 30 minutos (0.45), 45 minutos (0.1), 60 minutos (0.3)} o

Y(k) = {Participación Nula (0.1), Part. Escasa (0.5), Part. Media (0.3), Part. Alta (0.1)}.

Para clases de individuos, si k denota la provincia 'Álava', la variable Y = "Relación con la Actividad" puede ser especificada por:

–Y(k) = {Ocupado (0.47), Parado (0.11), Inactivo (0.42)}, que indica que el 47% de los individuos de Álava están ocupados, el 11% parados, etc..

Además, podemos encontrarnos con estudios que no estén basados en resultados experimentales o de encuestación únicos (Bock y Diday, 2000), sino que tienen en cuenta alguna inexactitud. De aquí surgen otro tipo de objetos simbólicos basados en resultados imprecisos: los datos probabilísticos o posibilísticos, los datos difusos, o los intervalos.

Los intervalos, a su vez, pueden resultar de dos fuentes: de observaciones o directamente de conocimiento experto. En el caso de datos resultantes de observaciones o medidas hay, por una parte, intervalos debidos a conocimiento impreciso y, por otra, intervalos debidos a variabilidad.

Vemos, pues, que existen diversas situaciones en las que la asociación de un único valor a un único individuo o a una única clase de individuos resulta en extremo limitada y no representa satisfactoriamente los contextos reales, más complejos. En estas circunstancias, los objetos simbólidos se muestran necesarios.

Métodos

La utilización de Objetos Simbólicos fue propuesta por E. Diday, y ha obtenido su máximo desarrollo en el marco del proyecto europeo SODAS ("Symbolic Official Data Analysis System"). (Ver Anexo I).

El proceso de creación de objetos simbólicos tiene como punto de partida consultas a una base de datos relacional. Por medio de estas consultas se extraen automáticamente grupos de individuos con características comunes, como, por ejemplo, familias, regiones, etc. Es decir, cada objeto simbólico puede describir un grupo o una clase de individuos.

Los objetos simbólicos creados son también almacenados en tablas, llamadas tablas simbólicas. Cada celda de estas tablas, con objetos simbólicos por filas y variables por columnas, puede contener datos de diferentes tipo, tales como:

–Un valor cuantitativo: edad (w) = 23;

–Un valor cualitativo: sexo (w) = mujer;

–Varios valores: (cuantitativos) peso (w) = {48, 52, 56} que significa que el peso de w puede ser 48 ó 52 ó 56; (cualitativos) estado civil (w) = {soltero, casado};

–Intervalo: edad (w) = [20, 25] que significa que la edad de w varía entre 20 y 25;

–Varios valores con pesos: edad (w) = [20 (0.65), 25 (0.35)], que puede ser un histograma o una función de pertenencia;

Siendo edad, sexo, estado civil y peso variables y w unidades.

Tabla 1: *Tabla Simbólica.*

	Sexo	**Edad**	**Profesión**
OS 1	{mujer(0.33), varón(0.67)}	{[25:57]}	{Técnico y profesionales superiores (0.35), Personal Directivo (0.25), Jefes Administrativos (0.4)}
OS 2	{mujer(0.5), varón(0.5)}	{[18:42]}	{Comerciantes y Vendedores (0.55), Administrativos (0.45)}

En la tabla 1, cada objeto simbólico (por filas) representa un grupo de individuos con características comunes descritas por 3 variables. La tabla entera corresponde a una consulta a la base de datos de la encuesta Población en Relación con la Actividad (P.R.A.). Así, por ejemplo, el objeto simbólico OS 2 representa a una clase de individuos formada por un 50% de mujeres y un 50% de varones, con edades comprendidas entre 18 y 42 años y que se dedican al comercio o venta (55%) o a la administración (45%).

Las variables que describen los objetos simbólicos pueden ser a su vez:

–*Variables con dominio Taxonómico*: Si ofrecen la posibilidad de definir una jerarquía en los valores que toma la variable. Esta taxonomía representa un conocimiento a priori de los datos.

Estado civil = soltero, no soltero (casado, viudo, divorciado/separado).

–*Variables Madre-Hija (o Dependencias Jerárquicas):* Si ofrecen la posibilidad de definir variables que no son aplicables a todos los individuos, pero sí lo son a individuos que verifican algunas propiedades.

SI Relación con la Actividad = parado
ENTONCES Tipo de Contrato es no aplicable.

–*Variables con Dependencias Lógicas (o Reglas):* Si ofrecen la posibilidad de definir conocimiento a priori de los datos en forma de restricción de las posibles combinaciones de valores para diferentes variables.

SI edad>65 ENTONCES Situación profesional = Retirado

Construcción de Objetos Simbólicos

Una base de datos relacional sigue una estructura de tabla donde cada registro representa un individuo. Una manera de obtener y describir información almacenada en bases de datos relacionales es mediante la construcción de objetos simbólicos. Éstos son creados agregando individuos en clases y describiendo propiedades de estas clases.

En el proceso de selección de la población se tienen en cuenta datos de varias tablas relacionadas, además de conocimiento adicional tal como taxonomías, variables madre-hija, etc...Los pasos de este proceso son:

–Consulta SQL en el módulo DB2SO del software SODAS que especifique qué datos relevantes tienen que ser procesados y qué atributos tienen que ser devueltos.

El formato general de una consulta de este tipo es:

```
SELECT id, atributo de grupo, resto de variables, [peso muestral]
FROM tabla
WHERE restricciones;
```

Estas consultas constan de un identificador único para cada individuo, una variable que agrupa a los individuos (atributo de grupo), otras variables que muestran la composición del grupo y, opcionalmente, una variable de peso muestral. La composición del grupo puede darse en porcentaje o en efectivos.

El resultado de una consulta, es decir, un grupo de registros, se considera la población en estudio. Si esta población es grande, incluso, se puede realizar un muestreo aleatorio.

–Descripción de cada grupo mediante objetos simbólicos para análisis futuros de estos grupos.

–Análisis estadísticos. Los análisis utilizados normalmente en cualquier tipo de población u elemento objeto de estudio estadístico —estadística descriptiva, métodos de clasificación....— pueden ser aplicados a estos objetos simbólicos.

Tipos de Objetos Simbólicos según su Construcción

–Según el nivel de agregación:

Los Objetos Simbólicos pueden describir individuos tanto como clases de individuos. Según esto, podemos distinguir objetos simbólicos de:

PRIMER ORDEN: Se dice que los objetos simbólicos son de primer orden cuando los datos se refieren a individuos. Por ejemplo, la variable Y = "Edad" para cada alumno k de un colegio:

$Y(k) = \{11\}$ o $Y(k) = [4, 13]$

SEGUNDO ORDEN: Se dice que los objetos simbólicos son de segundo orden (objetos agregados) cuando los datos se refieren a clases de individuos más o menos homogéneos. Como no todos los individuos de la misma clase toman el mismo valor en cada variable, habrá varias categorías que se aplicarán simultáneamente a la clase, normalmente con porcentajes especificados.

Ahora k denota una clase de individuos como puede ser un curso concreto del colegio del ejemplo anterior y la variable Y = "Edad" puede ser especificada por:

$Y(k) = \{10\ (0.2), 11(0.6), 12(0.2)\}$, que indica que el 20% de los individuos de ese curso tienen 10 años, el 60% tienen 11 años, etc..

Objetos de orden más alto pueden ser definidos de manera análoga mediante agregación sucesiva.

–Según el número de variables clasificadoras:

Las clases de individuos pueden crearse, aparte de con conocimiento experto o análisis previos para formar grupos, mediante una sola variable o combinación de varias. De esta forma, tenemos:

ATRIBUTO DE GRUPO SIMPLE: la formación de los grupos se hace mediante una sola variable. Se obtendrán tantos objetos simbólicos como modalidades tenga esa variable.

```
SELECT id, estado_civil, sexo, nivel_educación, edad,
       relación_actividad,...  FROM tabla
```

Con esta consulta se van a obtener 4 objetos simbólicos que describen el estado civil de la población: "Soltero", "Casado", "Viudo", "Divorciado/Separado", y que van a estar descritos por el resto de variables: sexo, edad, relación con la actividad...

```
os "Casado"(1684) =
   [sexo = {"Mujer"(0.498812), "Varón"(0.501188)}]
  ^[nivil = {"Estudios Secundarios"(0.224466), "Estudios
Universitarios"(0.100356), "Estudios Primarios o menos"(0.675178)}]
  ^[pral = {"Parados que han trabajado"(0.0647268),
"Ocupados"(0.465558), "Inactivos"(0.466746), "Parados que buscan
empleo"(0.00296912)}]
  ^[eden = {"35 a 44 años"(0.214964), "25 a 34 años"(0.0890736),
"16 a 24 años"(0.00653207), "55 a 64 años"(0.209026), "45 a 54
años"(0.269002), "65 y más años"(0.211401)}]
os "Soltero"(1001) =
   [sexo = {"Mujer"(0.487512), "Varón"(0.512488)}]
  ^[nivil = {"Estudios Secundarios"(0.522478), "Estudios
Universitarios"(0.232767), "Estudios Primarios o menos"(0.244755)}]
  ^[pral = {"Parados que han trabajado"(0.111888),
"Ocupados"(0.443556), "Inactivos"(0.370629), "Parados que buscan
empleo"(0.0739261)}]
  ^[eden = {"35 a 44 años"(0.0639361), "25 a 34 años"(0.328671),
"16 a 24 años"(0.491508), "55 a 64 años"(0.038961), "45 a 54
años"(0.043956), "65 y más años"(0.032967)}] ...
```

. ATRIBUTO DE GRUPO COMPUESTO: Si el atributo de grupo se compone del cruce de dos o más variables nominales. Se obtendrán tantos objetos simbólicos como producto de modalidades de las variables.

```
SELECT id, estado_civil & sexo, edad, relación_actividad,...
FROM tabla
```

En este caso se van a obtener 8 (4*2) objetos simbólicos combinación de las modalidades de estado civil con las de sexo: "Soltero Varón", "Soltero Mujer", "Casado Varón", "Casado Mujer", "Viudo Varón", "Viudo Mujer", "Divorciado/Separado Varón", "Divorciado/Separado Mujer".

```
os "Casado / Varón"(844) =
   [nivil = {"Estudios Primarios o menos"(0.622043), "Estudios
Secundarios"(0.25979), "Estudios Universitarios"(0.118167)}]
  ^[pral = {"Inactivos"(0.287075), "Parados que buscan
empleo"(0.00130055), "Ocupados"(0.677499), "Parados que han
trabajado"(0.034125)}]
  ^[eden = {"55 a 64 años"(0.188175), "16 a 24 años"(0.00517072),
"65 y más años"(0.205099), "45 a 54 años"(0.257138), "25 a 34
años"(0.0798497), "35 a 44 años"(0.264568)}]
os "Casado / Mujer"(840) =
   [nivil = {"Estudios Primarios o menos"(0.689833), "Estudios
Secundarios"(0.219648), "Estudios Universitarios"(0.0905192)}]
  ^[pral = {"Inactivos"(0.575217), "Parados que buscan
empleo"(0.00525723), "Ocupados"(0.312245), "Parados que han
trabajado"(0.10728)}]
  ^[eden = {"55 a 64 años"(0.176855), "16 a 24 años"(0.00878641),
"65 y más años"(0.178333), "45 a 54 años"(0.241147), "35 a 44
años"(0.261524), "25 a 34 años"(0.133355)}]
```

Además, en estas consultas pueden incluirse restricciones tanto en la variable que agrupa como en el resto.

Visualización

La visualización de un objeto simbólico se hace mediante un gráfico llamado Zoom Star. Esta representación está basada en los diagramas de Kiviat donde cada eje representa una variable. En el mismo gráfico pueden representarse variables categóricas, intervalo, con pesos, taxonomías,... sin sobrecargar el gráfico.

Hay dos tipos de representación, en 2D y 3D, que nos muestran diferentes niveles de detalle. La representación en 2D permite una impresión global del objeto, mientras que la representación en 3D nos da información más detallada.

En 2D los ejes están unidos por una línea que conecta los valores más frecuentes de cada variable. Si hubiera un empate del valor más frecuente en varias modalidades, la línea uniría las dos. Cuando existe una variable intervalo la línea se une a los límites mínimo y máximo y el área entera se rellena.

Como ejemplo, se han definido objetos simbólicos que sean grupos de población definidos por sexo, edad, estado civil, relación con la actividad y nivel de educación en la encuesta P.R.A (Población en Relación con la Actividad). Se han obtenido 314 objetos simbólicos, que son los cruces de todas las modalidades de estas variables.

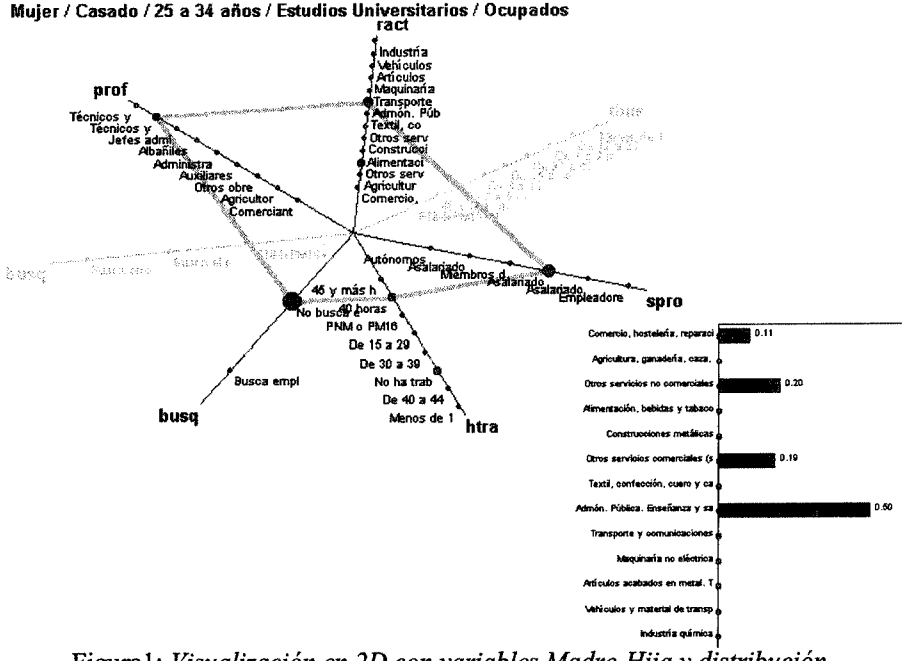

Figura1: *Visualización en 2D con variables Madre-Hija y distribución asociada de uno de los ejes (rama de actividad)*

En el gráfico de la figura 1 se pueden ver ejemplos de variables madre-hija. Las variables hijas que toman el valor N.A. (no aplicable) aparecen en el gráfico como un eje gris desactivado. A la derecha de éste se puede ver la distribución de una de las variables.

En la representación 3D, que se reproduce en la figura 2, se ve la distribución correspondiente a cada variable con pesos. Las variables numéricas se representan por rectángulos que van del valor mínimo al máximo.

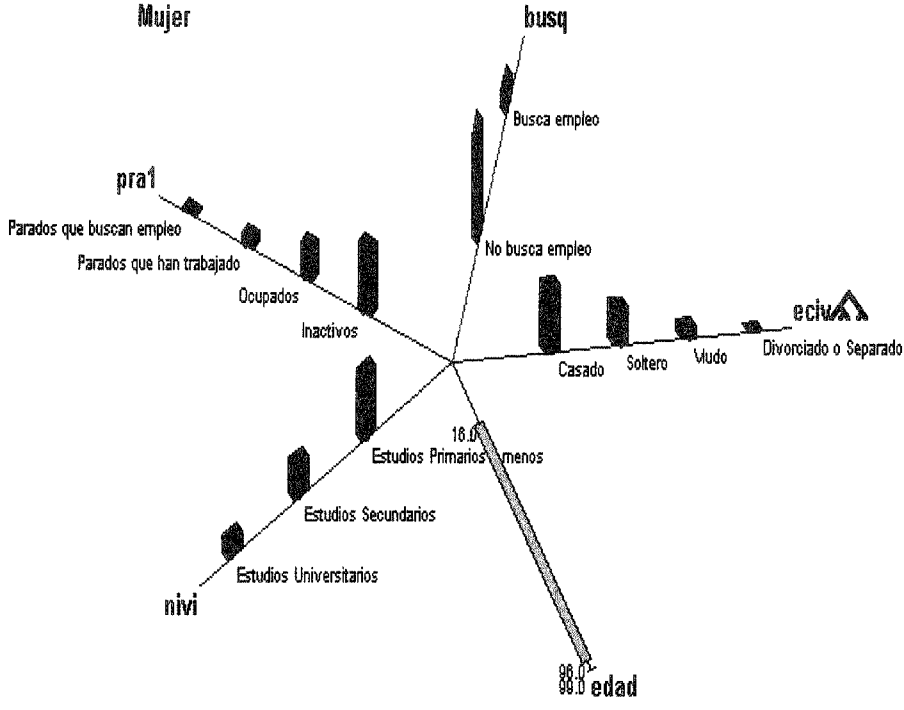

Figura 2: *Visualización en 3D con variable intervalo y taxonomía en la variable* eciv.

Ejemplos sobre encuestas de EUSTAT

En este apartado se toman dos ejemplos de utilización de objetos simbólicos en dos encuestas de EUSTAT: los datos de accidentes de tráfico proporcionados por la Dirección de Tráfico del Departamento de Interior del Gobierno Vasco y la Encuesta Económica de Servicios a Empresas.

Accidentes de tráfico 1999

Los datos aquí utilizados pertenecen a los accidentes de tráfico recogidos en un atestado por diversas causas. En concreto, porque las personas implicadas en ellos han presentado algún tipo de lesión o incluso han fallecido. Los datos hacen referencia a los atestados de 1999 recogidos por la Ertzantza en la Comunidad Autónoma de Euskadi.

A partir de esta información se han creado 18 objetos simbólicos en función de los factores concurrentes al accidente y en función de su gravedad. Las nueve modalidades que corresponden a los factores concurrentes son: distracción, infracción-velocidad, alcohol-drogas, mal estado del vehículo, mal estado de la vía, meteorología adversa, cansancio-enfermedad, inexperiencia, otros no definidos.

En cuanto a la gravedad del accidente se diferencian dos modalidades atendiendo a si ha habido muertos o no. Ello supone dos modalidades a cruzar con las nueve anteriores.

Estos objetos simbólicos podían haber sido descritos por otras variables, pero en este caso interesaba destacar cómo la situación que rodea un accidente y su gravedad pueden estar influidas por una serie de factores que aportan nueva información sobre las causas posibles del accidente. El siguiente bloque de texto especifica la definición del objeto simbólico 16:

```
"alcohol-drogas con muertos"(16) =
[epoca = {"primavera"(0.0625), "invierno"(0.25), "verano"(0.4375),
"otoño"(0.25)}]
^[facatmo = {"otros fact.atmos."(0.0625), "con lluvia"(0.0625), "buen
tiempo"(0.875)}]    ^[hora = {"7-9 horas"(0.1875), "16-18"(0.125),
"0-6 horas"(0.0625), "21-23"(0.25), "10-12 horas"(0.25),
"19-20"(0.125)}] ^ [intersec = {"no en intersec."(1)}] ^ [lumino =
{"pleno día"(0.5625), "noche ilum.suf."(0.125), "noche
ilum.insuf."(0.1875), "noche sin ilumi."(0.125)}] ^ [superf = {"seca
y limpia"(0.875), "mojada"(0.125)}] ^ [tipoacc = {"salida de
calzada"(0.1875), "frontal"(0.375), "vuelco"(0.0625),
"atropello"(0.125), "choque con obstáculo"(0.125), "otro"(0.125)}]
^ [tipodia = {"laborable"(0.25), "víspera de festivo"(0.375),
"festivo"(0.375)}]
^[tipovia = {"autopista o autovía"(0.25), "vía convencional"(0.5),
"resto vías"(0.25)}]
^[zona = {"variante"(0.0625), "zona urbana"(0.0625),
"carretera"(0.875)}]
```

En los gráficos Zoom Star de las figuras 3 y 4, vemos la descripción de alguno de los objetos simbólicos creados y las variables que intervienen en esos objetos (ejes axiales).

El trabajo realizado hasta ahora nos provee de una descripción de los objetos simbólicos no exenta de interés. Debemos mencionar, sin embargo, que existen

otros procedimientos que permiten profundizar y dar prioridad a las distintas modalidades de las variables que intervienen.

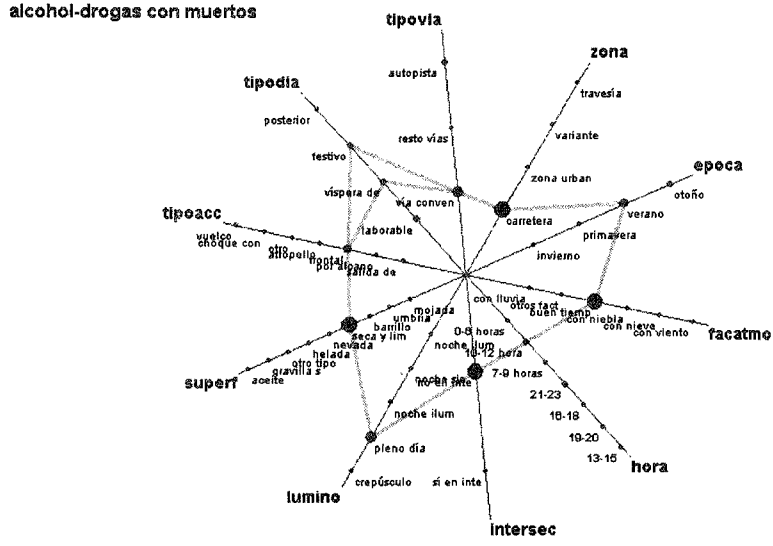

Figura 3: Zoom Star del objeto simbólico "alcohol-drogas con muertos".

Figura 4: Zoom Star de objeto simbólico "infracción-velocidad con muertos".

Así, mediante *la clasificación divisiva*, procedimiento DIV, se pueden encontrar grupos de accidentes en los que las variables que los definen tienen un comportamiento semejante, ordenando en una jerarquía cuáles son las variables que van discriminando.

Dicho método es un método de clasificación jerárquica "de arriba abajo" que persigue la detección y construcción de grupos homogéneos de objetos de población. Es decir, se empieza con un gran cluster de todos los objetos y se divide sucesivamente cada uno en otros más pequeños hasta que una regla de parada impide futuras divisiones.

Las variables que definen los accidentes son: época del año, factores atmosféricos, hora, intersección o no, luminosidad, superficie, tipo de accidente, tipo de día, tipo de vía, y zona.

El árbol que se crea forma ramas y cada una de ellas es doble. En un lado se sitúa la modalidad de corte y en el otro todas las demás modalidades que componen esa variable. Por ejemplo, con la variable factores atmosféricos tenemos en una rama los objetos ocurridos con lluvia, que es la modalidad discriminante, en la otra rama, el resto de los factores atmosféricos.

El árbol resultante es como sigue:

```
Explicated inertia : 50.905735
THE CLUSTERING TREE :
   - the number noted at each node indicates
     the order of the divisions
   - Ng <-> yes and Nd <-> no
(Nodo izquierdo <-> sí y Nodo derecho <-> no)
           +----  Classe 1 (Ng=1)
           !
      !----2- [epoca <= invierno]
      !    !
      !    !    +----  Classe 3 (Ng=2)
      !    !    !
      !    !----3- [hora <= 21-23]
      !         !
      !         +----  Classe 4 (Nd=1)
      !
 !----1- [facatmo <= con lluvia]
      !
      !    +----  Classe 2 (Ng=2)
      !    !
      !----4- [tipoacc <= salida de calzada]
           !
           +----  Classe 5 (Nd=12)
```

Las variables que cortan el árbol son:

–Los factores atmosféricos: con lluvia o sin lluvia.

–La época del año: invierno o en otra estación.

–La hora del día: de 9 a 11 de la noche u otra hora distinta al intervalo mencionado.

–Y, por último, el tipo de accidente: salida de calzada u otro tipo de accidente.

Las clases 1, 3 y 4 cuelgan de una rama del árbol tienen en común que son accidentes con lluvia:

–La clase 1 hace referencia a los accidentes que han tenido lugar con lluvia y en invierno.

–La clase 3 recoge a los accidentes con lluvia, fuera del invierno, pero a una hora más tardía, entre las 9 y las 11 de la noche.

–La clase 4 agrupa a los accidentes con lluvia, en otra época que no es invierno y fuera también del intervalo comprendido entre las 9 y las 11 de la noche. Se incluyen en esta clase los accidentes debidos a la inexperiencia y con resultado de muertos.

De la otra rama del árbol cuelgan las clases 2 y 5, son accidentes sin lluvia:

–La clase 2 se centra en accidentes sin lluvia y con salida de la calzada, unido al cansancio, sueño o enfermedad.

–La clase 5 representa a los accidentes sin lluvia y otro tipo de accidentes que no sean salidas de la calzada: choques, alcances o atropellos.

Las particiones contienen estos objetos simbólicos:

```
PARTITION IN 5 CLUSTERS:
Cluster 1 (n=1):
"meteorología adversa con muertos"
Cluster 2 (n=2):
"cansancio-enfermedad Sin muertos" "cansancio-enfermedad con muertos"
Cluster 3 (n=2):
"meteorología adversa Sin muertos" "mal estado vía con muertos"
Cluster 4 (n=1):
"inexperiencia con muertos"
Cluster 5 (n=12):
"otros Sin muertos" "distracción Sin muertos" "infracción-velocidad
Sin muertos" "alcohol-drogas Sin muertos" "mal estado vía Sin muertos"
"alcohol-drogas con muertos" "infracción-velocidad con muertos" "dis-
tracción con muertos" "mal estado vehículo Sin muertos" "inexperiencia
Sin muertos" "otros con muertos" "mal estado vehículo con muertos"
```

Servicios a empresas 1998

Objetivo

El fin de este apartado es el realizar una aplicación práctica del análisis de objetos simbólicos a través del software SODAS en el campo de las estadísticas económicas. Esta aplicación pone de manifiesto su utilidad en el campo de la estadística oficial, facilitando la tarea del análisis de datos. Tiene, por tanto, una doble intención; una es la de poner de manifiesto la utilidad del Análisis Simbólico de Datos, y otra, la de realizar un análisis de interés económico sobre un sector en plena fase de expansión. Lógicamente, tanto uno como otro objetivo, solo se podrán cumplir a un nivel ilustrativo, lo que supone desarrollos en futuros trabajos.

El sector servicios a empresas

El sector sobre el que va a centrar la aplicación es el de los servicios a empresas en el ámbito de la Comunidad Autónoma de Euskadi (en adelante CAE), y que hace referencia a las divisiones 70.2, 71, 72, 73 y 74[1] de la Clasificación Nacional de Actividades Económicas de 1993 (en adelante CNAE 1993).

Desde mediados de la década pasada, este sector ha mantenido una expansión continuada en la CAE, tal y como lo muestran los datos de la Tabla 2. Además, dentro de la denominación de Servicios a empresas, se esconden actividades muy diversas y de naturaleza muy diferente entre sí. Prueba de ello, es la alta dispersión respecto a la media obtenida para el conjunto del sector en las variables que serán objeto del estudio y que aparecen en el Tabla 3.

De ahí que este sector presente unas características muy apropiadas para la aplicación del análisis de objetos simbólicos y de las técnicas que brinda el programa SODAS. Así, por tanto, este trabajo se dirige a estudiar la homogeneidad y heterogeneidad entre las distintas actividades incluidas en los Servicios a empresas a través de sus ratios económicas. El trabajo, además, tiene interés en varios sentidos: en un sentido económico, para lograr un mayor conocimiento de los crecimientos experimentados por el sector, y en un sentido estadístico, por su aplicación en el campo de la elevación de datos y selecciones muestrales.

Tabla 2: *Participación de los Servicios a Empresas en el Valor Añadido Bruto Total.*

Año	%sobre el VAB total de la economía
1980	11.63
1985	11.63
1990	12.46
1995	14.38
1998	15.32

Fuente: Eustat Tablas Input-Output. Se incluyen los alquileres imputados de la vivienda propia.

Metodología utilizada

Los datos de partida para la realización de esta aplicación han sido los de la Encuesta de Servicios a empresas 1998 elaborados por EUSTAT cada dos años. Se han tomado como variables cuantitativas: el personal ocupado, la ratio por persona de producción, los consumos intermedios, el valor añadido[2], el excedente de explota-

[1] Se incluyen actividades que pertenecen a la división 90 de la CNAE93, ya que se incorporan en la realización de la Encuesta de Servicios a empresa que realiza EUSTAT.

[2] Valor Añadido Bruto a salida de fábrica (VAB): Viene definido por la diferencia entre la producción y los consumos intermedios.

ción[3] y las remuneraciones de personal, y variables cualitativas sobre el equipamiento de los respectivos establecimientos.

Se utilizó la actividad económica a 5 dígitos de la CNAE para la creación de los diferentes objetos simbólicos, punto de partida de este tipo de análisis.

Una vez obtenidos los objetos simbólicos (56) con sus variables correspondientes, se aplicó el método de Clustering Divisivo. Los análisis se realizaron para 10 clusters ampliándose posteriormente hasta 14. Se obtuvo una inercia muy alta: de 99,8. Se podría haber continuado la división pero el número de clusters se hacía ya muy grande para el análisis propuesto.

Dichos clusters fueron definidos a su vez como objetos simbólicos y sobre ellos se aplicaron algunos de los análisis de carácter más general que permite SODAS. Estos son: la edición de objetos simbólicos, en el módulo SOEditor, que permite la visualización de los mismos a través de tablas de datos, gráficos, así como la aplicación del módulo STAT para el cálculo de estadísticas básicas.

Resultados obtenidos

Clusters obtenidos en 1998

Como consecuencia del análisis de clusters, se obtuvieron las siguientes agrupaciones[4] de objetos de actividades (CNAE a 5 dígitos) de Servicios a empresas. Para cada uno de estos clusters, se calculó la estadística básica (media y CV). Esta información aparece ordenada de mayor a menor ratio de productividad en la Tabla 3 (*a* y *b*).

Los dos primeros clusters que aparecen en el análisis se destacan muy notablemente del resto, tanto por sus ratios de productividad en el primer caso como el de los consumos intermedios en el segundo.

–Cluster 2. Promoción inmobiliaria de viviendas (7111).

–Cluster 3. Gestión de soportes publicitarios (74402).

El siguiente grupo de clusters que se configuran en el análisis vienen marcados por *sus altos consumos intermedios*, (superiores a 35,504 millones por persona) y de entre ellos se puede destacar un grado medio alto de producciones por persona que podemos ordenar de la siguiente manera:

–Cluster 11. Otra producción inmobiliaria (70112).

–Cluster 6. Otras agencias de publicidad (74401).

–Cluster 7. Gestión de Sociedades de Cartera (74150).

–Cluster 4. Servicios Técnicos de Ingeniería y Asesorías de dirección y gestión empresarial (74141,74202).

---

[3] Excedente Bruto de explotación : Viene definido por la diferencia entre el VAB a coste los factores (deducidos los impuestos y incluidas la subvenciones) y los costes de personal.

[4] Se ha mantenido la numeración de clusters que facilitaba SODAS con el fin de utilizar los gráficos y tablas del mismo

El siguiente tipo de actividades viene marcado por las ratios de *producción* (19,697 millones/persona) que calificamos como *medios* entre los que se pueden distinguir, según el nivel de consumos, las siguientes agrupaciones:

Clusters con un nivel alto de consumos intermedios:

–Cluster 8. Actividades de Contabilidad, Diseño no industrial, Organización de ferias y Relaciones públicas.

Clusters con un *nivel medio de consumos intermedios* (entre 24,492 y 8,495 millones por persona):

–Cluster 14. Alquiler de otra maquinaria.

–Cluster 9. Mantenimiento de equipo informático, Consulta de aplicaciones informáticas, Consultoras legales, Servicios técnicos de arquitectos, Actividades anexas de distribución publicitaria y Limpiezas públicas.

Tabla 3a: *Denominación conglomerados y códigos de actividades.*

clusters	denominación	cnae
2	Promoción inmobiliaria	70111
3	Soportes publicitarios	74402
11	Otra promoción inmobiliaria	70112
6	Agencias publicitarias	74401
7	Gestión de sd. De cartera	74150
4	Servicios técnicos de ingeniería y Asesorías de dirección	74141, 74202
8	Actividades de contabilidad, Diseño no industrial, Organización de ferias y Relaciones públicas	74120, 74841, 74842, 74142
14	Alquiler de otra maquinaria	71340
9	Mantenimiento informático, Consulta de aplicaciones informáticas, Consultores legales, Servicios técnicos de arquitectos y Anexos a distribución publicitaria	72500, 72200,74111, 74201,74833, 90002
12	Otras actividades empresariales	74843
5	Otros ensayos y análisis técnicos	74302
10	Alquiler de medios de transporte, I+D, Notarías, Otros servicios técnicos	74112, 74204, 71320,71100, 71210,70202, 73200
15	APIs y Otra administración de bienes inmuebles, Alquiler de efectos personales, radio y TV, Actividades de Proceso de datos y Bases de datos relacionales y Resto de servicios a empresas	70310,70322, 71404,71401, 72400,72300, 74700,74812, 74831,74811, 74502,74203, 74113,74602, 74130,74301, 74501
1	Administración de inmuebles, Alquiler de vestuario, Consulta de equipo informático y otras actividades informáticas, Traductores, Detectives, Agencias de personal y actividades de empaquetado	70321,71402, 72100,72600, 74832,74820, 74601,74503

Clusters con una ratio media-baja por persona de consumos intermedios (8,495 millones):

–Cluster 12. Otras actividades empresariales.

–Cluster 5. Otros ensayos y análisis técnicos.

Por último, aparecen en el análisis los clusters con las productividades más bajas, pero que a su vez contemplan el mayor número de actividades de Servicios a empresas. Entre ellos, podemos distinguir los siguientes, ordenados de mayor a menor productividad por persona:

–Cluster 10. Alquiler de medios de transporte, I+D, Notarías, Otros servicios técnicos.

–Cluster 15. Agentes de la Propiedad Inmobiliaria y Otra administración de inmuebles, Alquiler de efectos personales, radio y TV, Actividades relacionadas con Proceso de Datos y Bases de Datos, Resto de actividades de servicios a empresas (Limpieza Industrial, Vigilancia, Estudios fotográficos...).

–Cluster 1. Administración de inmuebles, Alquiler de vestuario, Consulta de equipo informático y otras actividades informáticas, Traductores, Detectives, Agencias de personal, y Actividades de empaquetado.

Tabla 3b: *Principales Ratios Medios y de Dispersión por Cluster. Servicios a Empresas 1998. CAE. Miles de Pts.*

Cl.	producción/ persona		c.intermedios /persona		vab/ persona		remunerac. /persona		excedente brto/persona		personal medio		%/vab total sector
	media	CV	media	CV	media	CV	media	CV	media	CV	media	CV	
2	655.48	56	533.460	57	118543	54	6682	49	108544	58	8	50	17.8
3	244202	56	231945	57	11621	32	3923	52	5837	45	6	47	0.2
11	107729	40	86406	46	29422	39	5033	19	22215	44	9	44	1.3
6	102307	56	94994	57	12824	51	3754	52	8902	57	18	54	2.7
7	53505	56	36578	56	26163	146	21904	47	13571	282	500	57	0.8
4	50476	53	44743	52	12483	53	4969	58	8385	75	114	70	10.5
8	34667	55	28617	59	6566	49	3468	58	3520	60	34	85	6.3
14	34175	44	17425	48	23387	44	3999	25	7985	56	14	49	2
9	24175	52	14872	59	10168	54	3933	63	6393	73	110	34	24.9
12	31174	57	4731	57	25462	67	4219	55	20188	87	46	56	4.6
5	20728	47	3873	45	18386	50	2392	53	13687	59	28	54	1.1
10	17103	39	7660	74	11319	49	4050	50	7207	77	15	80	6.1
15	9905	39	4803	60	5154	36	3136	45	1982	73	2	29	15.3
1	4680	37	1206	50	3668	36	2812	49	1088	95	56	141	6.4

CV: Coeficiente de Variación de Pearson.

Para terminar este apartado podríamos concluir diciendo que en el sector Servicios a empresas de la CAE tienen una mayor representación las actividades de pro-

ductividad media y baja[5]. Sin embargo, destaca especialmente el peso que mantiene alguna de las actividades con altos ratios de productividad como es la Promoción Inmobiliaria (17,8% sobre el total del VAB de sector).

Comparación con los clusters en 1996

Comparando los resultados del análisis divisivo de 1996 y de 1998, se puede destacar que, el análisis de 1996 arroja prácticamente los mismos clusters que el de 1998, al margen de la variación en los valores de las variables de corte entre unas agrupaciones y otras, así como de las diferencias en la clasificación de algunas actividades. Esto aparece muy claramente para las actividades con mayores y menores productividades, y es, en las agrupaciones con valores intermedios, donde se manifiestan diferencias. Éste es el caso, por ejemplo, de las actividades de los clusters 9, 8 y 5 que conforman diferentes agrupaciones entre sí para el año 1996.

Una aplicación: Fusión de Encuestas

Una de las múltiples propiedades de los objetos simbólicos es la unión de objetos simbólicos que describen el mismo concepto. A partir de esta propiedad y debido al creciente interés por la Fusión de Encuestas, EUSTAT ha desarrollado una nueva aplicación para SODAS: la fusión de encuestas mediante objetos simbólicos.

La idea se basa en el *Statistical Matching* (Cox y Boruch, 1988; Winkler, 1995) donde un registro de un fichero A se uniría a uno o más de B mediante características comunes a ambos ficheros. Los registros "casados" de B no tienen porque pertenecer al mismo propietario del registro de A.

Esta nueva fusión se diferencia de la tradicional en que en lugar de unir individuo a individuo por variables comunes se fusiona por objetos simbólicos que describan el mismo grupo.

Antes de pasar a las aplicaciones, vamos a explicar la propiedad de Unión de Objetos Simbólicos (Bock y Diday, 2000) de la que también surgió la idea de Fusión mediante objetos simbólicos:

Se consideran dos matrices de objetos simbólicos diferentes, pero que describan lo mismo. Por ejemplo, una matriz que describa regiones mediante las variables de viviendas, y otra matriz con las mismas regiones, pero esta vez descritas por variables de empleo de individuos.

Sean X_1 y X_2 dos matrices de datos simbólicos con individuos correspondientes a los conjuntos E_1 y E_2 respectivamente, y con variables $Y_{11},..., Y_{ip}$ y $Y_{21},...Y_{2q}$ respectivamente. La unión de X_1 y X_2 se denota por ***unión*(X_1, X_2)** y es una matriz de datos simbólicos definidos:

[5] Siempre dentro de las ratios de la propia CAE y del propio sector estudiado sin comparar con otros sectores y con otras economías.

1. $E = E_1 \cap E_2$ (el conjunto de entidades u objetos simbólicos de la matriz simbólica resultante es la intersección de los dos conjuntos de entidades donde X_1 y X_2 están basados).
2. Las variables que describen *unión*(X_1, X_2) son $Y_{11},..., Y_{ip}, Y_{21},...Y_{2q}$ (la concatenación de las variables que describen X_1 y X_2).
3. Para cada $u \in E$ se define *unión*(X_1, X_2)(u):= (X_1(u), X_2(u)). La matriz de datos resultante X = *unión*(X_1, X_2) tiene el formato |E| x (p+q).
4. Las posibles taxonomías definidas en algunas variables de X_1 o X_2 se mantienen en *unión*(X_1, X_2).
5. Las posibles variables madre-hija definidas por reglas en X_1 o X_2 se mantienen en *unión*(X_1, X_2).

Las entidades que estén en X_1 (resp. X_2), pero no en X_2 (resp. X_1) se pierden en *unión*(X_1, X_2).

La fusión nos permite relacionar encuestas independientes con algún ítem en común. Estos items suelen ser las variables socio-demográficas (sexo, estado civil, edad, nivel de educación, relación con la actividad,...) presentes en todas las encuestas demográficas. Los grupos u objetos simbólicos se definen como el producto cartesiano de las modalidades de esas variables comunes. Éstos se crean por separado en cada encuesta para su unión posterior. El resultado del proceso son objetos simbólicos que resumen información de los dos ficheros de datos.

Se ha utilizado esta nueva técnica para la fusión de la encuesta de Presupuestos de Tiempo (EPT) con la de Condiciones de Vida (ECV), debido a que tienen unas variables comunes (las socio-demográficas) y es previsible que haya una relación entre ambas.

El primer paso sería definir las variables socio-demográficas comunes y crear objetos simbólicos para cada encuesta separadamente. El atributo de grupo para esos objetos sería el producto cartesiano de las variables comunes.

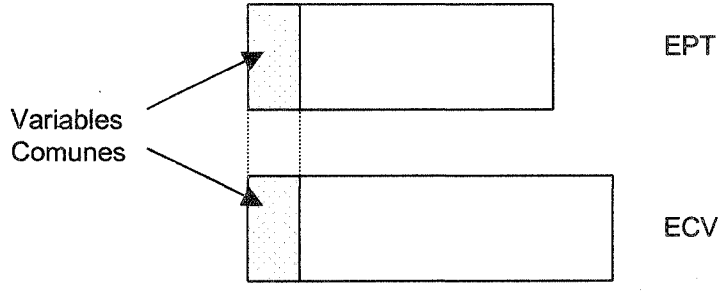

Figura 5: *Parte común y parte específica de dos encuestas independientes.*

Las variables comunes elegidas para este estudio fueron: sexo, estado civil, edad, relación con la actividad y nivel de educación.

El segundo paso sería unir objetos que describan el mismo colectivo. Así, para un mismo grupo tendríamos la descripción de las variables específicas de cada encuesta.

Como ejemplo, consideramos las siguientes matrices de datos:

Caso1:

X_1 es una matriz de datos simbólicos que describe grupos socio-demográficos por las siguientes variables de Presupuestos de Tiempo:

Y_{11}(limp) = participación en la limpieza

Y_{12}(prpc) = participación en preparación de comidas

Y_{13}(prac) = participación en la práctica de deportes

Y_{14}(cuip) = tiempo utilizado en el cuidado personal

Uno de los objetos de la matriz es:

```
os "Mujer Casado < 35 años Ocupado Media"(54) =
[limp = {"No Part."(0.347273), "Part.Escasa"(0.188186),
"Part.Media"(0.346782), "Part.Alta"(0.117759)}]
^[prpc = {"No Part."(0.0719004), "Part.Escasa"(0.400066),
"Part.Media"(0.436589), "Part.Alta"(0.0914451)}]
^[prac = {"No Part."(0.877218), "Part.Escasa"(0.122782)}]
^[cuip = [0:170]]
```

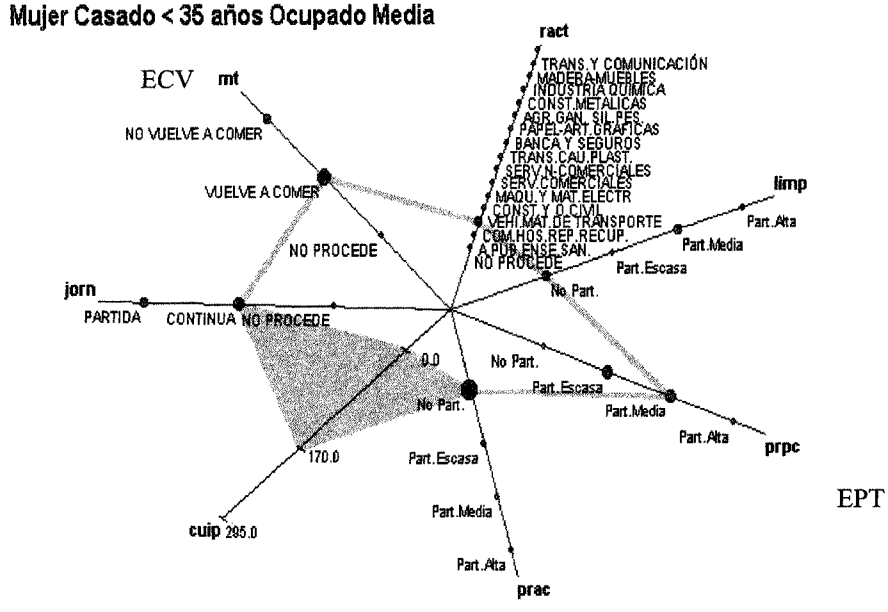

Figura 6: *Zoom Star del Objeto Fusión en 2D.*

Caso 2:

X_2 es una matriz de datos simbólicos que describe los mismos grupos socio-demográficos por las siguientes variables de Condiciones de Vida:

Y_{21} (jorna) = tipo de jornada laboral

Y_{22} (comt) = vuelta a casa para comer

Y_{23} (distr) = distancia al centro de trabajo

Y_{24} (ractp) = rama de actividad económica

```
os "Mujer Casado < 35 años Ocupado Media"(34) =
[jorna = {"Partida"(0.394297), "Continua"(0.434047),
"No Procede"(0.171656)}]
^[comt = {"Vuelve a comer"(0.637714), "No vuelve a comer"(0.345755),
"No Procede"(0.0165312)}]
^[ractp = {"Maqu.y Mat.Electr"(0.0497522), "Trans.y
Comunicación"(0.0317708), "Agr.Gan.Sil.Pes."(0.0201623),
"Com.Hos.Rep.Recup."(0.337456), "Industria Quimica"(0.0331782),
"Madera-Muebles"(0.0215493), "Serv.N-Comerciales"(0.078488),
"A.Pub.Ense.San."(0.137167), "Const.Metálicas"(0.0246369),
"Trans.Cau.Plast."(0.0199348), "Papel-Art.Gráficas"(0.0165312),
"Const.y O.Civil"(0.0235133), "Serv.Comerciales"(0.140833),
"Vehi.Mat.de Transporte"(0.0167604), "Banca y Seguros"(0.048266)}]
```

El resultado de la fusión se utilizaría como método alternativo de imputación. Es decir, a los objetos de la primera encuesta se les imputaría los valores que toman en la segunda y viceversa.

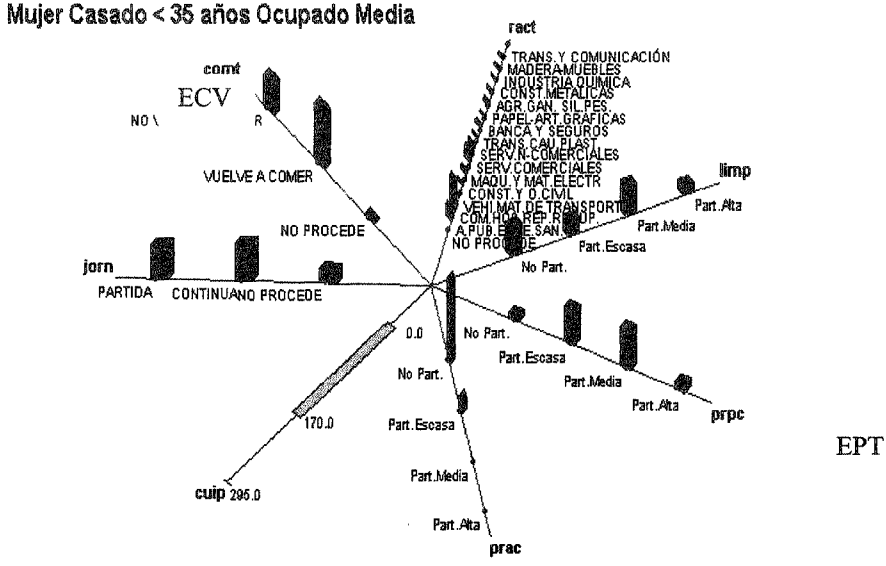

Figura 7: *Zoom Star del Objeto Fusión en 3D.*

Es decir, el colectivo "mujeres casadas <35 años ocupadas y con estudios medios" están ocupadas mayoritariamente en comercio, hostelería, servicios comerciales y administración pública. Además, tienen poca participación en la limpieza y algo más en la preparación de comidas.

Conclusiones

- Como principal conclusión hay que indicar las ventajas que ofrece la utilización de los Objetos Simbólicos como unidades estadísticas, en el campo de la Estadística Oficial y que son, entre otras, las siguientes:
  1. Resumir un gran volumen de datos en unidades estadísticas altamente significativas, para aplicar sobre las mismas los distintos análisis estadísticos.
  2. Proporcionar una manera eficiente de presentación y visualización de datos con estructura compleja de cara a su difusión.
  3. Facilitar la transformación de los datos en estructura de matriz.
- El Software SODAS permite la definición, creación, visualización y análisis de los Objetos Simbólicos de forma fácil para el productor estadístico y para el analista, permitiendo el enriquecimiento del análisis de los datos con un reducido coste añadido.
- Este tipo de análisis permite la fusión de encuestas lo que añade un aprovechamiento de la información recogida desde dos planos:
  1. Por un lado, tenemos una parte común en varias encuestas, por ejemplo, variables sociodemográficas, que se recogen habitualmente como variables independientes. Entonces se puede añadir información procedente de variables de otras encuestas e interpretarla en conjunto. Los cuestionarios así pueden ser más reducidos y conlleva un ahorro en la recogida de información o un aprovechamiento de los recursos.
  2. Por otro lado, esta fusión de encuestas permite que tengamos para ciertos niveles territoriales mayor información por lo que se podrá hablar de ámbitos territoriales menores.

Orientaciones bibliográficas sobre objetos simbólicos

Además de los textos citados en el apartado de referencias, tienen interés en el campo de los objetos simbólicos los siguientes documentos:

Calvo, P. (2000). Aplicaciones de los Objetos Simbólicos en la Estadística Oficial. Cuaderno Técnico de EUSTAT.

Diday, E. (1992). *Analyse des données et classification automatique numérique et symbolique.* Seminario Internacional de Estadística en Euskadi. Volumen 27. EUSTAT.

Diday E. (1993). From data to knowledge, boolean, probabilist and belief objects for symbolic data analysis. Une introduction a l'analyse des donnes symboliques.

INRIA-Rocquencourt. Domaine de Voluceau. Le Chesnay Cedex. Tutorial at IFCS'93.

Diday, E. (1998). *Symbolic Data Analysis: a theory and tool for Data Mining*. Invited conference at IFCS'98. Roma. Springer Verlag.

Diday, E. y Hebrail, G. (1998). *Symbolic Data Analysis: some in and out.* KESDA'98.

Iztueta, A. y Calvo, P. (1999). Uses of Symbolic Objects in Official Statistics. ISI'99.

Noirhomme-Fraiture, M. y Rouard, M. (1998). *Representation of Sub-populations and Correlation with Zoom Star*. Proc. NTTS'98 Sorrento, Italia.

Informes Internos del Proyecto SODAS (1996-99):

- Diday E. *Extracting Information from very Extensive Data Sets by Symbolic Data Analysis*. University Paris 9 Dauphine.
- Hebrail G., Lechevallier Y. y Stephan V*. (1997) *SODAS-RDBMS Interface.*
- Stéphan V*. (1997) *Extracting Symbolic Objects from Relational Databases.*
- Stéphan, V., Hebrail, G. y Lechevallier, Y*. (1997) *Building Symbolic Objects from Relational* Databases.
- Stéphan V., Hebrail G., Lechevallier Y*. (1997) *Improving Symbolic Descriptions of sets of Individuals: the reduction of assertions.*

Referencias

Bock, H. H. y Diday, E. (2000). Analysis of Symbolic Data. Springer-Verlag.

Cox, L.H. y Boruch R.F. (1988). *Record Linkage, Privacy and Statistical Policy.* Journal of Official Statistics. Vol. 4, No. 1, pp. 3-16.

Winkler W.E. (1995) Matching and Record Linkage. En B.G. Cox y col. *Business Survey Methods*, New York: John Wiley, 355-384.

Anexo I

"*SODAS: Symbolic Official Data Analysis System*" es el proyecto nº 20281 de la Comisión Europea, Directorio General III, Industrial RTD, EUROSTAT, programa DOSES.

En este proyecto intervienen varios miembros pertenecientes a Universidades, Empresas, Institutos de Estadística Oficial y Centros de Investigación de la Unión Europea, entre los que se encuentra EUSTAT en calidad de Estadística Oficial.

La primera etapa del Proyecto SODAS ha finalizado con el resultado del software SODAS 1.04. En este software están implementados los módulos que se han ido nombrando en el artículo, así como varios más. Una nueva etapa del proyecto está en marcha, llamada ASSO, que mejorará lo alcanzado hasta ahora e incluirá nuevos módulos para el manejo y el análisis de Objetos Simbólicos.

Información sobre el Proyecto puede encontrarse en las siguientes direcciones:
http://www-rocq.inria.fr/sodas/wp1/welcome.htm
http://www.ucm.es/info/otri/complutecno/ind_info.htm ---> SODAS 1.04
Además, el propio software SODAS 1.04 puede bajarse de la página:
http://www.cisia.com/download