

**APLICACIONES DE LOS OBJETOS SIMBOLICOS
EN LA ESTADISTICA OFICIAL**

Patricia Calvo Garrido



**EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA**

Donostia-San Sebastián, 1
01010 VITORIA-GASTEIZ
Tel.: 945 01 75 00
Fax.: 945 01 75 01
E-mail: eustat@eustat.es
www.eustat.es

APLICACIONES DE LOS OBJETOS SIMBÓLICOS EN LA ESTADÍSTICA OFICIAL

Patricia CALVO GARRIDO

Asistente Técnico Estadístico en EUSTAT

RESUMEN

En este trabajo se aborda principalmente la definición, creación y visualización de Objetos Simbólicos a partir de encuestas almacenadas en Bases de Datos Relacionales. Además, se ha dedicado un capítulo a la estadística básica de estos Objetos, dejando análisis más elaborados para siguientes cuadernos técnicos. En cada paso, se presentan las ventajas de este nuevo análisis frente a un análisis clásico de las mismas.

La utilización de los objetos simbólicos en la Estadística Oficial proporciona capacidad para describir datos con estructura compleja. La información sobre esta estructura va a ser extraída de la base de datos y va a ser descrita mediante objetos simbólicos. Estos objetos pueden describir individuos o clases de individuos.

La estructura del cuaderno va a ser la siguiente:

En la *Introducción* se presentan el proyecto SODAS, a partir del cual ha surgido este cuaderno, las nociones básicas de definición de un objeto simbólico y cuál es el origen de esta nueva notación.

El *Capítulo Segundo* describe la extracción de conocimiento de una base de datos mediante consultas o queries a la misma y cómo el resultado de esas consultas es transformado en objetos simbólicos.

En el *Tercer Capítulo* se expone la forma de visualizar objetos simbólicos mediante tablas, gráficos o el lenguaje exclusivo de objetos simbólicos.

El *Cuarto Capítulo* se centra en la estadística descriptiva de objetos simbólicos

Finalmente, se presentan las conclusiones del trabajo y el futuro de esta nueva teoría dentro de la estadística oficial.

PALABRAS CLAVE:

Objeto simbólico, aserción, objeto booleano, objeto probabilístico, base de datos relacional, consulta.

Indice

INDICE	3
INTRODUCCIÓN	4
INTRODUCCIÓN Y OBJETIVOS	4
DESCRIPCIÓN DEL PROYECTO	4
NECESIDAD DE OBJETOS SIMBÓLICOS.....	5
TABLAS DE DATOS	6
DEFINICIÓN FORMAL DE OBJETO SIMBÓLICO.....	7
Tipos de Objetos Simbólicos	8
Objetos Simbólicos Modales	8
EJEMPLO	9
OBJETOS SIMBÓLICOS DE PRIMER Y SEGUNDO ORDEN.....	9
OBJETOS SIMBÓLICOS COMO MEDIDA DE IMPRECISIÓN.....	10
PROCEDENCIA DE LOS OBJETOS SIMBÓLICOS	11
OBTENCIÓN DE OBJETOS SIMBÓLICOS A PARTIR DE CONSULTAS A UNA BASE DE DATOS	12
CONSTRUCCIÓN DE ASERCIONES	12
CONSTRUCCIÓN DE VARIABLES MADRE-HIJA O DEPENDENCIAS JERÁRQUICAS.....	14
ADICIÓN DE TAXONOMÍAS EN LOS DOMINIOS DE LAS VARIABLES	15
REFINAMIENTO DE ASERCIONES	16
UNIÓN DE ASERCIONES	18
VENTAJAS DEL USO DE OBJETOS SIMBÓLICOS.....	20
VISUALIZACIÓN DE OBJETOS SIMBÓLICOS. ZOOM STAR.....	21
VISUALIZACIÓN EN UNA TABLA SIMBÓLICA.....	21
VISUALIZACIÓN EN ESTRELLAS.....	22
VISUALIZACIÓN COMO SOL.....	24
COMPARACIÓN DE VARIOS OBJETOS SIMBÓLICOS.....	25
ANÁLISIS DE LA EVOLUCIÓN DE UN OBJETO SIMBÓLICO.....	27
VENTAJAS DEL USO DE OBJETOS SIMBÓLICOS.....	29
ESTADÍSTICA BÁSICA DE OBJETOS SIMBÓLICOS	30
FRECUENCIAS DE VARIABLES MULTINOMIALES	30
FRECUENCIAS DE VARIABLES INTERVALO.....	31
BILOT	32

CAPACIDADES PARA VARIABLES PROBABILÍSTICAS.....	32
CONCLUSIONES	33
OTROS USOS DE LOS OBJETOS SIMBÓLICOS	33
FUTURO DEL PROYECTO.....	34
BIBLIOGRAFÍA	35

Introducción

Introducción y objetivos

Este cuaderno tiene como objetivo la iniciación al Análisis Simbólico de Datos en la Estadística Oficial. Para ello, se ha aplicado este nuevo análisis al estudio de varias encuestas de EUSTAT para conocer sus ventajas frente a un análisis de datos clásico.

Los métodos clásicos de análisis estadístico estaban diseñados para una situación relativamente simple. Primero, los datos eran obtenidos para individuos simples usando encuestas, experimentos, archivos, etc. Segundo, las variables estaban bien definidas y tercero, esas variables tenían un solo valor o categoría en cada individuo.

A veces, el mundo real es demasiado complejo para ser descrito por estos modelos relativamente simples. Para poder tratar estos casos, se introduce el concepto y varios tipos de objetos simbólicos.

Un Objeto Simbólico es un modo de representación de datos complejos que surge al analizar grandes ficheros de datos. En los Institutos de Estadística una de las tareas más importantes es resumir esos ficheros en otros más reducidos con nuevas unidades estadísticas perdiendo la menor información posible. Esas nuevas unidades estadísticas van a ser los objetos simbólicos que van a extender el Análisis de Datos Clásico al correspondiente Análisis Simbólico de Datos.

Descripción del Proyecto

La utilización de Objetos Simbólicos propuesto por E. Diday, ha obtenido su máximo desarrollo en el marco del proyecto europeo SODAS.

“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.

El objetivo del proyecto SODAS es facilitar el uso de Análisis Simbólico de Datos tanto en la estadística oficial como en compañías, y por consiguiente demostrar que esas nuevas técnicas cumplen varias necesidades de usuario:

- Análisis de datos con estructura compleja.
- Mejores explicaciones de resultados estadísticos.

- Representación, manipulación y análisis de conceptos y metadata.
- Intercambio de datos entre miembros comunitarios de Estadísticas Oficiales.

Este objetivo se alcanzará mediante:

- El desarrollo de un software para probar y evaluar el procesamiento de datos simbólicos.
- La construcción de un sistema orientado a conceptos para estadística oficial.

El nuevo software incluirá:

- Herramientas genéricas para almacenar, consultar y actualizar objetos simbólicos,
- Herramientas para adquirir objetos simbólicos de grandes bases de datos,
- Una colección de métodos de análisis de datos dedicados a objetos simbólicos: métodos univariantes descriptivos, clustering, construcción de árboles de decisión, discriminación, y análisis factorial,
- Herramientas para transformar objetos simbólicos en objetos "estándar" y después aplicar métodos estándar de análisis de datos,
- Herramientas ergonómicas para presentar al usuario los resultados de los métodos.

Necesidad de Objetos Simbólicos

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

Por ejemplo, para individuos, la variable $Y = \text{"Minutos dedicados a la práctica de deporte al día"}$ 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:

$$Y(k) = [20,60] \text{ o}$$

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

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

Para clases de individuos, si k denota la región 'Álava', la variable $Y = \text{"Relación con la Actividad"}$ puede ser especificada por:

$$Y(k) = \{\text{Ocupado } (0.47), \text{Parado } (0.11), \text{Inactivo } (0.42)\},$$

que significa que el 47% de los individuos de Álava están ocupados, el 11% parados, etc..

Tablas de Datos

La creación de Objetos Simbólicos tiene como punto de partida tablas de una Base de Datos Relacional donde,

- Existen varias tablas a distintos niveles relacionadas entre sí.
- Los datos y la metadata aparecen por separado por lo que no se repite información.
- Se pueden tener a la vez una tabla de datos y varias tablas de metadata en diferentes idiomas.

Los Objetos Simbólicos generados van a ser también almacenados en tablas. Estas tablas, con objetos simbólicos por filas y variables por columnas, serán el comienzo para los distintos algoritmos de Análisis Simbólico de datos. Cada celda de estas tablas puede contener datos de diferentes tipos, 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 o 52 o 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.

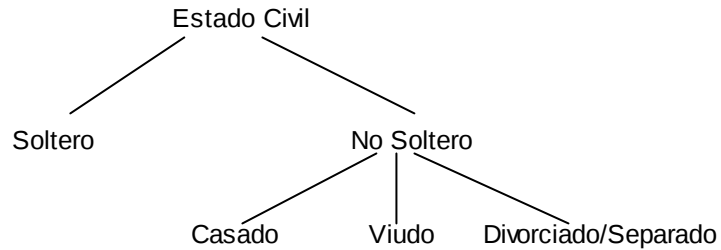
Es decir, los valores que toman los individuos en las variables pueden ser *no atómicos* (un grupo de valores, un intervalo de valores o una distribución de probabilidad).

Tabla1. Tabla de Objetos Simbólicos

	Sexo	Edad	Relación con la Actividad
OS 1	{mujer(0.33), varón(0.67)}	{[25:57]}	{ocupado(0.67), inactivo(0.33)}
OS 2	{mujer(0.25), varón(0.75)}	{[16:42]}	{ocupado(0.25), parado(0.25), inactivo(0.50)}
OS n	{mujer(0.5), varón(0.5)}	{[27:29]}	{ocupado(1)}

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

1. *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.



Esto mismo en forma de objeto simbólico:

`estado_civil= {soltero, no soltero = {casado, viudo, divorciado/separado}};`

2. *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 N.A. (no aplicable);

3. *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

Según qué tipo de dato contengan las celdas que componen los Objetos Simbólicos, estos pueden ser de varias clases:

- **Objetos Booleanos**: si las celdas asociadas son del tipo a), b), c) y d) descritas anteriormente.
- **Objetos Modales**: si al menos una celda de la fila correspondiente contiene pesos.

Así pues, el proceso de obtención de Objetos Simbólicos tiene como esquema:

Consultas a una Base de Datos → Creación de Objetos Simbólicos → Tablas de Objetos Simbólicos → Análisis Simbólico de Datos.

Definición Formal de Objeto Simbólico

Se define **Objeto Simbólico** como “una descripción que se expresa con la ayuda de un conjunto de sucesos (o propiedades) inducidos por los valores tomados por las variables” (véase [4]).

Una variable y es una función $\Omega \rightarrow V$ donde Ω es el conjunto de “objetos elementales” y O el conjunto de observación donde la variable toma sus valores ($V \subset O$).

Tipos de Objetos Simbólicos

- Sucesos elementales: $e_i = [y_i = V_i]$

Es una función $e_{y_i, V_i}: \Omega \rightarrow \{\text{verdadero, falso}\}$ tal que $e_{y_i, V_i}(w) = \text{verdadero}$ si y sólo si $y_i(w) \in V_i$.

Ej. $[longitud = 0.52] \text{ o } [color \in \{\text{rojo, azul}\}]$.

- Objeto Aserción: $a = [y'_1 = V_1] \dot{\cup} \dots \dot{\cup} [y'_q = V_q]$ donde $V_i \subset O_i$ está definido por la función $a_{y'} V: \Omega \rightarrow \{\text{verdadero, falso}\}$ tal que $a_{y'} V(w) = \text{verdadero}$ si y sólo si para todo $i=1, \dots, q$ se tiene $y_i(w) \in V_i$.

Es una conjunción de sucesos elementales que deben ser verdaderos simultáneamente para el mismo objeto elemental $w \in \Omega$.

Ej. $[longitud = 0.52] \wedge [color \in \{\text{rojo, azul}\}] \wedge [forma = \text{rectangular}]$.

- Objeto Horda: $h = [y'_1(u_1) = V_1] \dot{\cup} \dots \dot{\cup} [y'_p(u_p) = V_p]$ está definido por la función $h_y V: \Omega^q \rightarrow \{\text{verdadero, falso}\}$ tal que $\forall W = (w'_1, \dots, w'_q) \in \Omega^q$, $h_y V(W) = \text{verdadero}$ si y sólo si $\forall i$ $y'_i(w'_i) \in V_i$.

- Objeto de Síntesis: $s = h_1 \dot{\cup} \dots \dot{\cup} h_k$, es la conjunción de k objetos horda definidos respectivamente sobre cada uno de los conjuntos $H_1, \dots, H_k$ con $h_i \in H_i$.

Objetos Simbólicos Modales

Anteriormente definíamos un objeto modal como aquel en el que por lo menos una celda contiene valores con pesos. Ahora vamos a clasificar esos objetos según si los pesos (o modos) afectan a todo el objeto o sólo a sus valores.

Objetos modales externos: los modos afectan globalmente al suceso.

$a_x = \hat{1} M_i [y_i = V_i]$ donde x hace referencia a la semántica (posibilista, probabilista, ...).

Ej. : *frecuentemente* $[Edad = [16, 24]]$ (posibilista)

Objetos modales internos: los modos afectan a los valores tomados por las variables.

$a_x = \hat{1} [y_i = M_i V_i]$ donde x hace referencia a la semántica.

Ej. : $[Estado_civil = \text{no soltero } (0.7), \text{ soltero } (0.3)]$ (probabilista)

Definición de un Objeto en Intención y en Extensión

Definición de un objeto en intención: El objeto es descrito por las propiedades que lo caracterizan.

Definición de un objeto en extensión: La extensión de un objeto simbólico es el conjunto de objetos elementales de Ω que lo satisfacen. Se denota $|s|_\Omega$ o $\text{Ext}(s)$.

En el caso booleano, $\text{Ext}(s) = \{w \in \Omega / a(w) = \text{verdadero}\}$

En el caso modal, dado un umbral α , $\text{Ext}(s) = \{w \in \Omega / a(w) \geq \alpha\}$

En este cuaderno se utilizarán fundamentalmente sucesos elementales y aserciones modales internos y definidos en intención.

Ejemplo

Tenemos la siguiente tabla con 4 individuos (por filas) y 3 variables (por columnas):

	y_1 : Sexo	y_2 : Edad	y_3 : Educación
Ind1	Mujer	25	universitaria
Ind2	Mujer	60	primaria
Ind3	Hombre	38	secundaria
Ind4	Hombre	54	secundaria

El conjunto de "objetos elementales" es $\Omega = \{\text{Ind1}, \text{Ind2}, \text{Ind3}, \text{Ind4}\}$.

El conjunto de observación de la variable y_1 es $O_1 = \{\text{mujer}, \text{hombre}\}$ y de la misma forma para y_2 y y_3 .

Así, Ind1 se puede describir mediante la siguiente aserción booleana:

$\text{Ind1} = [\text{Sexo} = \text{mujer}] \wedge [\text{Edad} = 25] \wedge [\text{Educación} = \text{universitaria}]$

Para aclarar los términos de definición en intención y en extensión, se toma otro objeto simbólico cualquiera

$a = [\text{Sexo} = \text{hombre}] \wedge [\text{Educación} = \text{secundaria}]$.

Esta descripción sería una definición en intención del objeto a.

Su definición en extensión sería:

$\text{Ext}(a) = \{\text{Ind3}, \text{Ind4}\}$, puesto que los individuos 3 y 4 cumplen 'a'.

Objetos Simbólicos de Primer y Segundo Orden

Primer Orden

Se dice que los objetos simbólicos son de primer orden cuando los datos se refieren a individuos.

Sea $E = \Omega = \{1, \dots, n\}$ un universo de individuos (objetos elementales) .

Por ejemplo, la variable $Y = \text{"Edad"}$ 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 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.

Sea $E = \{C_1, C_2, \dots\}$ un sistema de clases $C_i \subseteq \Omega$ (objetos agregados).

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

$$Y(k) = \{10(0.2), 11(0.6), 12(0.2)\},$$

que significa 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 (agregar 2 cursos diferentes del mismo nivel y compararlos en base a la edad).

Objetos Simbólicos como Medida de Imprecisión

Podemos encontrar con estudios que no estén basados en resultados experimentales o de encuestación únicos, pero que tienen en cuenta alguna inexactitud. Es aquí donde surgen otro tipo de objetos simbólicos basados en resultados imprecisos. Esto incluye datos probabilísticos o posibilísticos, datos difusos, o intervalos.

Los intervalos 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: el resultado ξ_j de una observación o de una medida es un intervalo $[a_{ij}-\delta, a_{ij}+\delta]$ donde a_{ij} es el valor observado y δ caracteriza la imprecisión del instrumento de medida. Por otra parte, hay intervalos debidos a variabilidad: sea $a_{ij}^1, \dots, a_{ij}^k$ observaciones de la variable j para el objeto i . El resultado de resumir esas k observaciones es el intervalo $x_{ij} = [\underline{x}_{ij}, \overline{x}_{ij}]$ donde $\underline{x}_{ij}, \overline{x}_{ij}$ son el mínimo y el máximo valores observados, respectivamente.

Ejemplo

Consideremos las estimaciones mediante intervalos de confianza de una determinada medida en grupos definidos por territorio, relación con la actividad laboral y rama de actividad económica. Al no conocer la medida exacta, se va a definir un objeto simbólico que incluirá la medida intervalo.

$$Y = [\text{terr} = \text{Alava}] \wedge [\text{pra1} = \text{Ocupados}] \wedge [\text{ract2} = \text{Agricultura, ganadería y pesca}] \wedge [\text{estimación} = [32.75, 40.8]].$$

Es decir, la estimada de este colectivo estaría entre 32.75 y 40.8.

Un experto no está a menudo 100% seguro sobre sus afirmaciones y en ese caso expresa dudas, creencias, etc. Normalmente se usan intervalos para describir conocimiento experto incluida incertidumbre.

Procedencia de los Objetos Simbólicos

Los Objetos Simbólicos pueden ser el resultado de varios métodos:

- 1) De consultas a Bases de Datos Relacionales.
- 2) De Análisis de Datos de tablas estándar para formar grupos (análisis factorial, análisis de clusters,...).
- 3) De conocimiento experto.
- 4) De series temporales.
- 5) De datos confidenciales (para ocultar datos iniciales mediante una menor precisión).

En nuestro caso, utilizando varias encuestas se han generado objetos simbólicos mediante consultas directas a la base de datos.

En los capítulos siguientes se van a desarrollar diferentes tipos de objetos simbólicos haciendo especial hincapié en las ventajas de su utilización con respecto a los métodos tradicionales de análisis de datos.

Obtención de Objetos Simbólicos a partir de Consultas a una Base de Datos

La forma más directa de obtener objetos simbólicos es mediante consultas a una base de datos. Estas consultas extraen automáticamente grupos de individuos con características comunes, como por ejemplo familias, regiones, etc. En realidad es un proceso de generalización de un conjunto de datos clásico almacenados en una base de datos teniendo en cuenta relaciones entre diferentes tablas. Una vez creados los objetos simbólicos puede aplicárseles un proceso de especialización para reducir la super-generalización o unir varios aplicando el operador de unión.

Construcción de Aserciones

Una base de datos relacional sigue una estructura de tabla donde cada tupla representa un individuo.

Una manera de obtener y describir información almacenada en bases de datos relacionales es mediante la construcción de objetos simbólicos. Estos son creados agregando individuos en clases y describiendo las propiedades de esas 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 de 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 (*id*), una variable que agrupa a los individuos (*atributo de grupo*), otras variables que muestran la composición del grupo y opcionalmente un variable peso muestral. La composición del grupo puede darse en porcentaje o en efectivos.

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

- Descripción de cada grupo mediante aserciones, para análisis futuros de estos grupos.

Atributo de Grupo Simple

Si la formación de los grupos se hace mediante una sola variable. Se obtendrán tantos objetos simbólicos como modalidades tenga esa variable.

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.

Consultas con Restricciones

Se impone un filtro sobre los datos que va a devolver la consulta. La restricción puede ser sobre la variable que indica el grupo o sobre cualquier otra variable que describa el grupo.

Todas las consultas pueden partir de individuos simples o de grupos previamente definidos, como regiones, viviendas,...

Ejemplos

Atributo de Grupo Simple:

```
SELECT id, estado_civil, sexo, nivel_educación, edad, pra1,...
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, pra1...

```
os 'Casado'(1684) =
  [sexo = {"Mujer"(0.498812), "Varón"(0.501188)}]
  ^[nivi1 = {"Estudios Secundarios"(0.224466), "Estudios
Universitarios"(0.100356), "Estudios Primarios o menos"(0.675178)}]
  ^[pra1 = {"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)}]
  ^[nivi1 = {"Estudios Secundarios"(0.522478), "Estudios
Universitarios"(0.232767), "Estudios Primarios o menos"(0.244755)}]
  ^[pra1 = {"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:

```
SELECT id, estado_civil & sexo, edad, pra1,...
FROM tabla
```

En este caso vamos 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) =
  [nivi1 = {"Estudios Primarios o menos"(0.622043), "Estudios
Secundarios"(0.25979), "Estudios Universitarios"(0.118167)}]
  ^[pra1 = {"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) =
  [nivi1 = {"Estudios Primarios o menos"(0.689833), "Estudios
Secundarios"(0.219648), "Estudios Universitarios"(0.0905192)}]
  ^[pra1 = {"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)}]
```

Consulta con Restricciones:

```
SELECT id, estado_civil, sexo, pra1,...
FROM tabla
WHERE 25 < edad < 35
```

Se van a obtener 4 objetos simbólicos describiendo cada uno de los estados civiles como en el primer ejemplo, pero esta vez los individuos descritos van a estar solamente entre los 25 y 35 años.

```
os 'Casado'(150) =
  [sexo = {"Varón"(0.382255), "Mujer"(0.617745)}]
  ^[nivi1 = {"Estudios Primarios o menos"(0.418376), "Estudios
Secundarios"(0.442944), "Estudios Universitarios"(0.13868)}]
  ^[pra1 = {"Inactivos"(0.148825), "Parados que buscan empleo"(0.0059544),
"Parados que han trabajado"(0.179955), "Ocupados"(0.665266)}]

os 'Soltero'(329) =
  [sexo = {"Varón"(0.563178), "Mujer"(0.436822)}]
  ^[nivi1 = {"Estudios Primarios o menos"(0.189587), "Estudios
Secundarios"(0.517487), "Estudios Universitarios"(0.292927)}]
  ^[pra1 = {"Inactivos"(0.0509509), "Parados que buscan empleo"(0.0723981),
"Parados que han trabajado"(0.19166), "Ocupados"(0.684991)}]
```

Puede observarse que con la restricción impuesta sobre la edad el número de efectivos en cada objeto disminuye. Compárense las cifras entre paréntesis de los objetos del primer ejemplo y los de este.

Construcción de Variables Madre-Hija o Dependencias Jerárquicas

Como se vio en el capítulo 1, las variables madre-hija definen variables que no son aplicables a todos los individuos, pero sí lo son a individuos que verifican algunas propiedades.

Muchas de las variables que tratamos en las encuestas dependen de la respuesta de otras variables anteriores.

Por ejemplo, la variable busq2 (búsqueda del primer empleo u otro) en la Encuesta de Población en relación con la Actividad (P.R.A) sólo es válida para los individuos que han declarado buscar empleo. No tiene sentido aplicar esta variable a un individuo si en otra variable (busq1) se indica que no busca empleo. La regla en este caso sería,

SI busq1="no busca empleo" ENTONCES busq2=N.A. (No aplicable)

La creación de variables madre-hija transforma el *fichero de individuos*. La variable hija (en este caso busq2) va a depender ahora de los valores que tome la variable madre. Si la variable madre toma el valor donde la hija es aplicable (*busca empleo*), la hija tomará uno de sus valores posibles (*primer empleo, otro empleo,...*). Si embargo, si la madre toma valores donde la hija no es *aplicable* (*no busca empleo*), esta tomará el valor N.A.

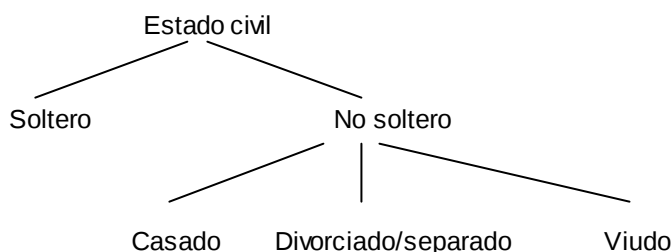
En el *fichero de aserciones* aparece la nueva regla. En este caso está enunciada en positivo, es decir, dónde es aplicable en lugar de dónde no es aplicable.

busq2 is applicable if busq1 in {"Busca empleo"}

Adición de Taxonomías en los dominios de las variables

Las variables taxonómicas permiten definir un orden en sus valores. La definición de la jerarquía de valores precisa de un conocimiento a priori.

Así, por ejemplo, se podría definir un orden en los valores de la variable *estado civil*:



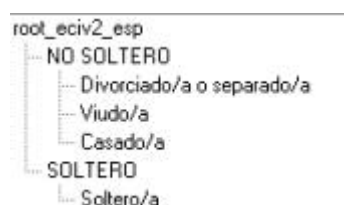
Anteriormente a la consulta SQL, hay que crear una tabla en la base de datos que defina la taxonomía.

Para el ejemplo anterior de la variable estado civil, se muestran la tabla creada en la Base de Datos y la visualización en el módulo DB2SO una vez creada la taxonomía.

Tabla en Base de Datos

Descripcion	Taxonomia
Casado/a	NO SOLTERO
Viudo/a	NO SOLTERO
Divorciado/a o separado/a	NO SOLTERO
Soltero/a	SOLTERO

Visualización de la taxonomía



En el *fichero de aserciones* aparece la descripción de la variable incluida la taxonomía.

```

variable eciv2
  nominale {"Soltero/a", "Casado/a", "Divorciado/a o separado/a", "Viudo/a",
    "SOLTERO", "NO SOLTERO","root_eciv2"}
  multiple, mode=probabilist;
  
```

```
taxonomy {
  "SOLTERO" = {"Soltero/a"},
  "NO SOLTERO" = {"Casado/a", "Divorciado/a o separado/a", "Viudo/a"},
  "root_eciv2" = {"SOLTERO", "NO SOLTERO"}
}
```

Refinamiento de Aserciones

Las aserciones anteriormente descritas, al considerar todos los individuos, incluso los atípicos, dan lugar a una super-generalización. Por eso, ahora se propone un método basado en el volumen que rechace individuos atípicos en la descripción final.

Para conseguir aserciones más homogéneas se recurre al refinamiento de las mismas. Este procedimiento hace decrecer el volumen de las aserciones quitando algunos individuos en cada grupo con un umbral mínimo de cobertura. Además, se consigue que el número de solapamientos entre aserciones sea menor.

Las aserciones se transforman en otras más específicas, más fáciles de interpretar y que dan una mejor descripción de las características en términos de homogeneidad. La calidad de las aserciones producidas es muy importante ya que van a ser el input para los métodos de análisis simbólico de datos.

Para realizar la reducción, se ha adaptado un criterio de volumen que mide un índice general en la unión de individuos:

$$vol/(a) = \prod_{i=1, \dots, p} card(d_i)$$

Este criterio de volumen no se puede aplicar con características numéricas y nominales a la vez. Para superar problemas de escala entre características numéricas y nominales, se transforman las características numéricas en ordinales. Esta codificación permite tener un criterio homogéneo entre todos los tipos de descripciones sin dar preferencia a ninguno.

Se codifica un dominio numérico buscando una distribución uniforme en cada intervalo de puntos. Es decir, en el caso numérico d_i es un conjunto de intervalos "uniformes", que generalizan las descripciones individuales. También se da más peso a las variables generales que a las taxonómicas para que la descripción sea más simple.

Se fija un umbral α , que es el poder mínimo de cobertura de la aserción. El poder de cobertura de una aserción a_C se calcula con, y la función de pertenencia correspondiente a a_C es:

$$Re\ c(a_C) = \frac{\sum_{w \in C} a(w)}{card(C)}$$

donde

$$a(w) = \begin{cases} 1 & \text{si } w \text{ pertenece a la descripción } a \\ 0 & \text{resto} \end{cases}$$

Cada paso del algoritmo consiste en quitar de la descripción valores que maximicen la reducción del hiper-cubo α con un α fijo.

Ejemplo

Se ha aplicado una reducción con un $\alpha=85\%$ al ejemplo siguiente:

```
os "Mujer 25 a 34 años Universitarios Ocupados" =
  [eciv2 = {"Casado"(0.24562), "Soltero"(0.75438)}]
  ^[prof2 = {"Administrativos"(0.0320081), "Auxiliares
  Administrativos"(0.166976), "Técnicos y Profesionales Superiores"(0.179049),
  "Técnicos y Profesionales Medios"(0.427945), "Personal Directivo"(0.0239909),
  "Otro personal de servicios"(0.0241), "Comerciantes y vendedores"(0.0999308),
  "Jefes Administrativos"(0.0459999)}]
  ^[spro1 = {"Asalariados de la Empresa Pública"(0.0778989),
  "Autónomos"(0.0258654), "Asalariados de la Administración Pública"(0.427499),
  "Asalariados del Sector Privado"(0.468736)}]
  ^[tjor = {"A tiempo completo (3 o más h/día)"(1)}]
  ^[tcon = {"PNM+PM16+Inactivos+No Asalariados y miembros de
  cooperativas"(0.0258654), "Indefinido fijo"(0.401042), "Otros"(0.0494191),
  "Temporal(formación, estacional, ocasional u otros)"(0.523673)}]
```

Después de la reducción, la aserción tiene 5 individuos menos considerados atípicos:

```
os 0.85 "Mujer 25 a 34 años Universitarios Ocupados" =
  [eciv2 = {"Casado"(0.227209), "Soltero"(0.772791)}]
  ^[prof2 = {"Auxiliares Administrativos"(0.191749), "Técnicos y
  Profesionales Superiores"(0.176412), "Técnicos y Profesionales
  Medios"(0.466286), "Otro personal de servicios"(0.0276755), "Comerciantes y
  vendedores"(0.0850538), "Jefes Administrativos"(0.0528245)}]
  ^[spro1 = {"Asalariados de la Administración Pública"(0.461722),
  "Asalariados del Sector Privado"(0.538278)}]
  ^[tjor = {"A tiempo completo (3 o más h/día)"(1)}]
  ^[tcon = {"Indefinido fijo"(0.423784), "Temporal(formación, estacional,
  ocasional u otros)"(0.576216)}]
```



Fig.1: Matriz de volumen antes (cuadros sin rellenar) y después del refinamiento. Los cuadros negros indican el grado de solapamiento entre aserciones.

Unión de Aserciones

Se consideran dos arrays de aserciones diferentes, pero que describan los mismos objetos. Por ejemplo, un array de aserciones que describa regiones mediante las variables de viviendas, y otro array con las mismas regiones pero esta vez descritas por variables de empleo de individuos.

Sea X_1 y X_2 dos arrays de datos simbólicos arbitrarios con individuos correspondientes a los conjuntos E_1 y E_2 respectivamente, y con variables $Y_{11}, \dots, Y_{ip}$ y $Y_{21}, \dots, Y_{2q}$ respectivamente. La unión de X_1 y X_2 se denota por **unión(X_1, X_2)** y es un array de datos simbólicos definidos:

1. $E = E_1 \cap E_2$ (el conjunto de entidades u objetos simbólicos del array simbólico 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 $Y_{11}, \dots, Y_{ip}, Y_{21}, \dots, Y_{2q}$ (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 $X = \text{unión}(X_1, X_2)$ tiene el formato $|E| \times (p+q)$.
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)**.

Usos de la Unión de Objetos en la Estadística Oficial: Fusión de Encuestas

Se ha encontrado una nueva aplicación de la unión de objetos simbólicos que consiste en la unión de aserciones provenientes de distintas encuestas. Esta unión de encuestas sirve para obtener información adicional, imputar datos, obtener conclusiones sobre causas y posibles efectos,...

La fusión de encuestas utilizando objetos simbólicos se diferencia de la fusión de encuestas tradicional en que en lugar de fusionar individuo a individuo por variables comunes se fusiona por objetos simbólicos cada uno describiendo un grupo.

La fusión nos permite relacionar encuestas independientes con algún ítem en común. En el marco del proyecto SODAS esta comparación sería entre encuestas de distintos países de la Unión Europea.

Ejemplo

Un ejemplo que ilustraría esta nueva utilización en EUSTAT sería 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.

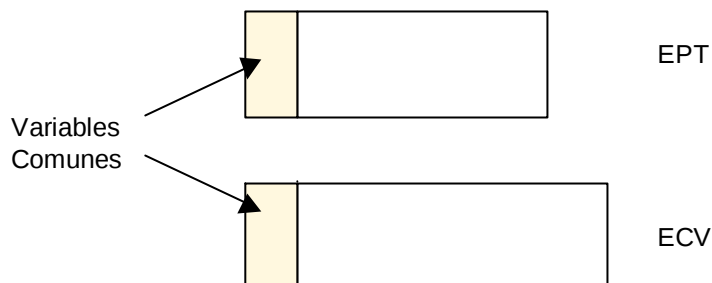


Fig.2: Parte común y parte específica de dos encuestas independientes.

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 aserciones que describan el mismo colectivo. Así, para un mismo grupo tendríamos la descripción de las variables específicas de cada encuesta.

Consideramos los siguientes arrays de datos:

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

$Y_{11}(\text{limp}) =$ Participación en la Limpieza

$Y_{12}(\text{prpc}) =$ Participación en Preparación de Comidas

$Y_{13}(\text{prac}) =$ Participación en la Práctica de Deportes

$Y_{14}(\text{cuip}) =$ Tiempo utilizado en el Cuidado Personal

Uno de los objetos del array es:

```
os 'Mujer Casado < 35 años Ocupado Media54) =
  [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]]
```

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

$Y_{21}(\text{jorna}) =$ Tipo de Jornada Laboral

$Y_{22}(\text{comt}) =$ 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 Medio' =
  [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.METALICAS"(0.0246369),
"TRANS.CAU.PLAST."(0.0199348), "PAPEL-ART.GRAFICAS"(0.0165312), "CONST.Y
O.CIVIL"(0.0235133), "SERV.COMERCIALES"(0.140833), "VEHI.MAT.DE
TRANSPORTE"(0.0167604), "BANCA Y SEGUROS"(0.048266)}]
```

El objeto simbólico unión es el siguiente:

```
os 'Mujer Casado < 35 años Ocupado Medio' =
  [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]]
  ^[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), "AGR.GAN. SIL.PES."(0.0201623),
"MADERA-MUEBLES"(0.0215493), "SERV.N-COMERCIALES"(0.078488), "TRANS.Y
COMUNICACIÓN"(0.0317708), "A.PUB.ENSE.SAN."(0.137167),
"CONST.METALICAS"(0.0246369), "COM.HOS.REP.RECUP."(0.337456), "INDUSTRIA
QUIMICA"(0.0331782), "TRANS.CAU.PLAST."(0.0199348), "PAPEL-
ART.GRAFICAS"(0.0165312), "CONST.Y O.CIVIL"(0.0235133),
"SERV.COMERCIALES"(0.140833), "VEHI.MAT.DE TRANSPORTE"(0.0167604), "BANCA Y
SEGUROS"(0.048266)}]
```

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.

Ventajas del uso de Objetos Simbólicos

Cada tupla resultado de una consulta a la base de datos se convierte en una nueva unidad estadística llamada objeto simbólico.

Con esta nueva unidad estadística se pueden realizar todo tipo de análisis estadísticos, con la ventaja de que tratamos grupos en lugar de individuos. Estos grupos pueden contener información de varias tablas relacionadas.

Visualización de Objetos Simbólicos. Zoom Star.

Los Objetos Simbólicos van a poder ser visualizados de tres formas diferentes:

- En una tabla simbólica,
- Mediante gráficos de estrellas y
- Mediante el lenguaje específico de objetos simbólicos, SOL (Symbolic Object Language).

Visualización en una Tabla Simbólica

En una tabla simbólica las filas son objetos simbólicos y las columnas variables. Según sea el tipo de variable en las celdas aparecerán distribuciones, intervalos...

En la siguiente tabla se tienen 4 objetos simbólicos que describen los 4 estados civiles mediante las variables sexo, relación con la actividad,...

	sexo	pra1
Casado	Varón (0.50), Mujer (0.50)	Inactivos (0.47), Ocupados (0.47), Parados que han trabajado (0.06), Parado
Soltero	Varón (0.51), Mujer (0.49)	Inactivos (0.37), Ocupados (0.44), Parados que han trabajado (0.11), Parado
Viudo	Varón (0.19), Mujer (0.81)	Inactivos (0.92), Ocupados (0.06), Parados que han trabaj
Divorciado o Separado	Varón (0.36), Mujer (0.64)	Inactivos (0.13), Ocupados (0.64), Parados que han trabaj

Sobre esta tabla se puede hacer todo tipo de selecciones sobre objetos simbólicos y sobre variables. En este caso, se ha seleccionado el objeto “Viudo” y la variable Relación con la actividad.

	sexo	pra1
Casado	Varón (0.50), Mujer (0.50)	Inactivos (0.47), Ocupados (0.47), Parados que han trabajado (0.06), Parado
Soltero	Varón (0.51), Mujer (0.49)	Inactivos (0.37), Ocupados (0.44), Parados que han trabajado (0.11), Parado
Viudo	Varón (0.19), Mujer (0.81)	Inactivos (0.92), Ocupados (0.06), Parados que han trabaj
Divorciado o Separado	Varón (0.36), Mujer (0.64)	Inactivos (0.13), Ocupados (0.64), Parados que han trabaj

Los identificadores de las categorías pueden ser reemplazados por identificadores generados automáticamente, y activar una ventana que contenga la metadata con las etiquetas completas.

	sexo	nvl	pnl
Casado	AB01 (0.50), AB02 (0.50)	AC01 (0.69), AC02 (0.22), AC03 (0.10)	AD01 (0.47), AD02 (0.47), AD03
Soltero	AB01 (0.51), AB02 (0.49)	AC01 (0.24), AC02 (0.52), AC03 (0.23)	AD01 (0.37), AD02 (0.44), AD03
Viudo	AB01 (0.19), AB02 (0.81)	AC01 (0.92), AC02 (0.06), AC03 (0.03)	AD01 (0.92), AD02 (0.06)
Divorciado o Separado	AB01 (0.36), AB02 (0.64)	AC01 (0.84), AC02 (0.13), AC03 (0.13)	AD01 (0.13), AD02 (0.84)

	Multiple Nominal Model	Description
1	AC01	Estudios Primarios o menos
2	AC02	Estudios Secundarios
3	AC03	Estudios Universitarios
4	Not Applicable	
5	Missing Value	

Visualización en Estrellas

Hay dos tipos de representación en estrella, en 2D y 3D, que nos proveen 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.

La representación en estrella 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.

La siguiente tabla resume la representación de cada variable dependiendo de su tipo.

<i>Tipo de Variable</i>	<i>Descripción del eje</i>
Cuantitativa	Eje graduado
Categórica	Puntos igualmente distribuidos en el eje
Categórica sin pesos	Eje dibujado en negro
Categórica con pesos	Eje dibujado en claro
No aplicable	Eje dibujado en gris

El límite de variables a representar es 24 y el de modalidades dentro de una variable es 15.

Haciendo doble-clic encima de un eje aparece solo la distribución de la variable elegida (histograma). Además se pueden visualizar las taxonomías y las dependencias de una variable haciendo clic en el icono que aparece debajo del eje correspondiente.

Los gráficos se pueden girar hacia la derecha, izquierda, arriba y abajo para su mejor visualización.

2D Zoom Star

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. Se han obtenido 314 objetos simbólicos, que son los cruces de todas las modalidades de estas variables.

En este caso al ser una encuesta, las probabilidades se han calculado teniendo en cuenta los pesos muestrales.

En el gráfico siguiente se pueden ver ejemplos de variables madre-hija. Las variables hijas que toman el valor N.A. aparecen en el gráfico como un eje gris desactivado. A la derecha de este se puede ver la distribución de una de las variables.

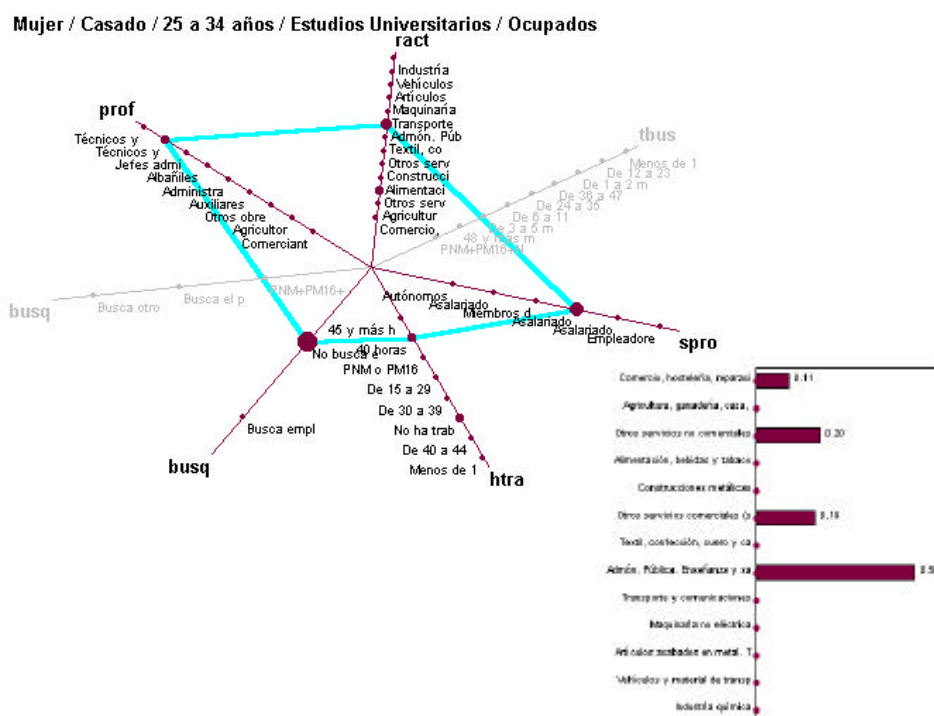


Fig.3: Visualización en 2D con variables Madre-Hija y distribución asociada de uno de los ejes (rama de actividad).

3D Zoom Star

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

Como ejemplo, la distribución del objeto simbólico "Mujer" en la encuesta P.R.A correspondiente a un trimestre en Álava.

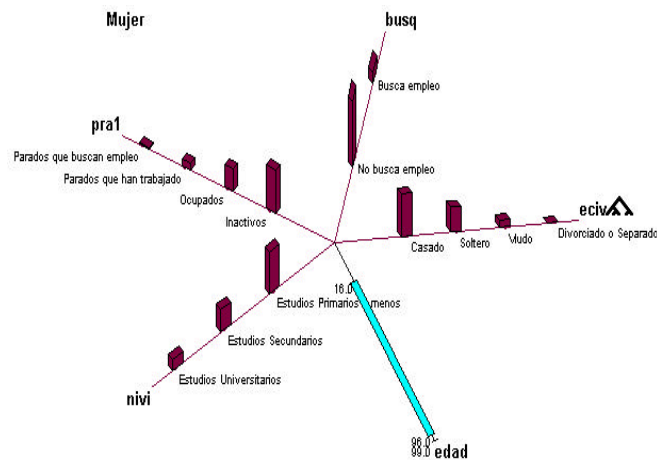


Fig.4: Visualización en 3D con variable intervalo y taxonomía en la variable eciv.

Visualización como SOL

El SOL es un lenguaje natural que puede ser leído (y escrito) fácilmente por los usuarios.

Definición de Variables

En el fichero de aserciones, primero se definen las variables que van a describir los objetos simbólicos. Se detalla de qué tipo son y qué modalidades tienen. A continuación se incluyen las reglas creadas entre las variables.

```
variable busq1
  nominale {"Busca empleo", "No busca empleo"}
  multiple, mode=probabilist;
```

```
variable busq2
  nominale {"Busca el primer empleo", "Busca otro empleo (no el primero)",
  "PNM+PM16+ No busca empleo"}
  multiple, mode=probabilist;
```

```
variable tbus1
  real [0:99]
  interval;
```

```
busq2 is applicable if busq1 in {"Busca empleo"}
tbus2 is applicable if busq1 in {"Busca empleo"}
```

Definición de Objetos Simbólicos

Después de la definición de las variables, se incluyen todos los objetos simbólicos en SOL.

Mujer / Casado / 25 a 34 años / Estudios Universitarios / Ocupados =

busq1 = No busca empleo (1.00)

And busq2 = Not Applicable

And prof2 = Comerciantes y vendedores (0.11), Auxiliares administrativos (0.19), Jefes administrativos (0.19), Técnicos y profesionales superiores (0.33), Técnicos y profesionales intermedios (0.19)

And ract2 = Comercio, hostelería, reparación (0.11), Otros servicios no comerciales (0.20), Otros servicios comerciales (0.19), Admón. Pública. Enseñanza y sanidad (0.50)

And tbus2 = Not Applicable

And spro1 = Autónomos (0.11), Asalariados del Sector Privado (0.19), Asalariados de la Administración Pública (0.62), Asalariados de la Empresa Pública (0.09)

And htra2 = 40 horas (0.34), De 15 a 29 horas (0.10), De 30 a 39 horas (0.19), No ha trabajado en la semana (0.28), De 40 a 44 horas (0.09)

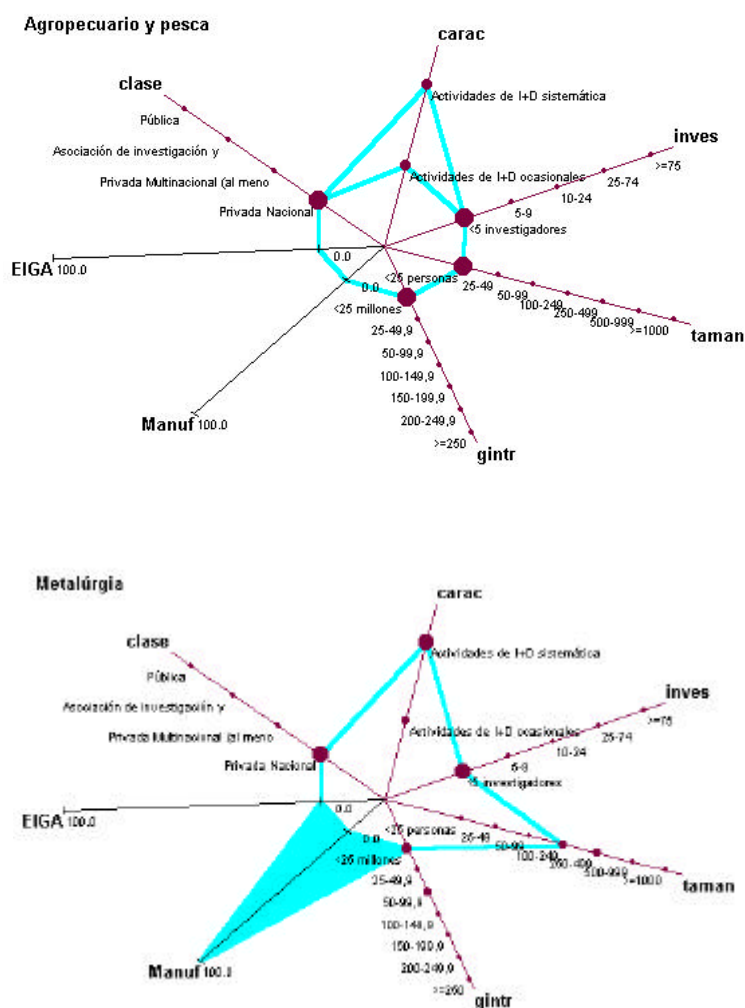
Comparación de Varios Objetos Simbólicos

La comparación de varios objetos simbólicos es más fácil utilizando la representación en 2D. Lo que se compararía es si la forma de las líneas poligonales que unen los ejes de las estrellas son similares.

Ejemplo

De la encuesta de Empresas que realizan I+D en el País Vasco, hemos construido objetos simbólicos que describen ramas de actividad económica. De las 18 ramas disponibles hemos elegido dos de ellas para su comparación, "Agropecuaria y Pesca" y "Metalurgia".

Las variables elegidas para describir las ramas son: Clase de la empresa, Carácter de la actividad I+D, Número de Investigadores en la actividad, Tamaño de la Empresa en personal, Gastos Intramuros, Porcentaje de la Investigación dedicado productos manufactureros y de energía.

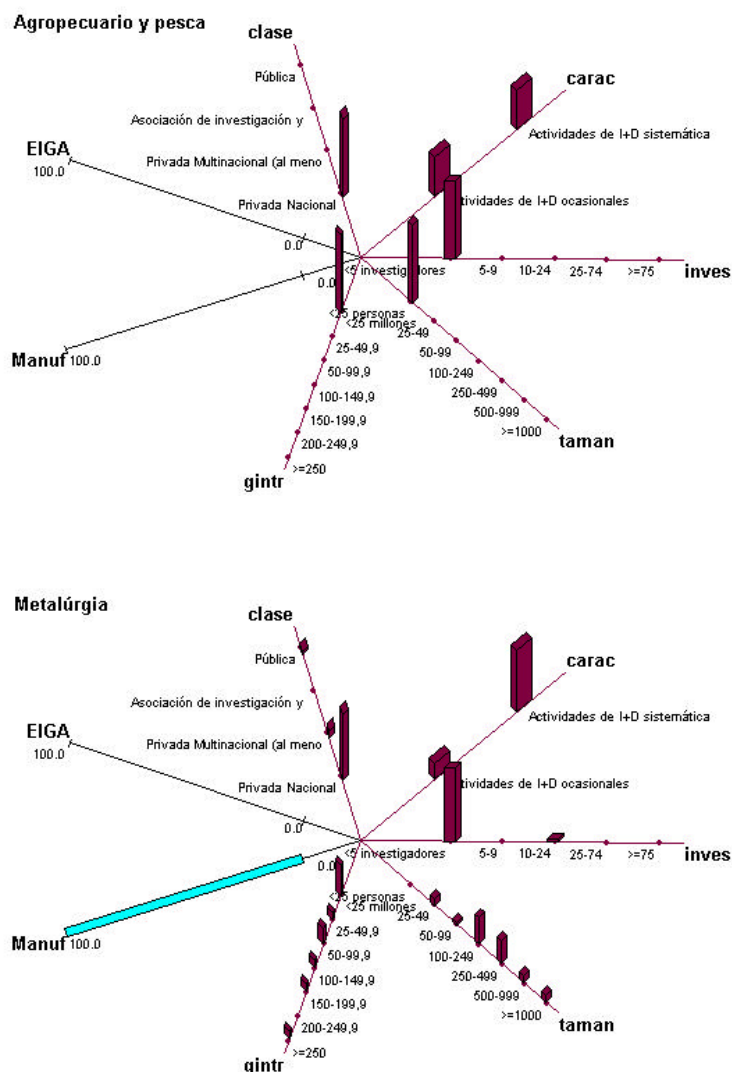


De los gráficos se sacan las siguientes conclusiones:

Las dos ramas se diferencian en el Carácter de las actividades de I+D, en "Agropecuaria y Pesca" las actividades se alternan entre sistemáticas y ocasionales, mientras que en "Metalúrgica" las actividades son mayoritariamente sistemáticas. Otra diferencia es el Tamaño de las Empresas, en la industria Metalúrgica son más grandes que en la Agropecuaria. Además, la industria Metalúrgica utiliza el 100% de los gastos intramuros en investigar productos manufactureros.

La comparación mediante histogramas (representación en 3D) también nos da información relevante sobre las distribuciones.

Representamos los gráficos en 3D de las mismas ramas de actividad económica que en el ejemplo anterior, para obtener más información.



Ahora pueden apreciarse mejor las diferencias en distribución de las dos ramas. En “Metalurgia” se puede ver que las distribuciones de las variables Tamaño de la Empresa (taman) y Gastos Intramuros (gintr) están mucho más dispersas entre las modalidades, mientras que en “Agropecuario y pesca” las distribuciones de estas dos variables se centran únicamente en un solo valor.

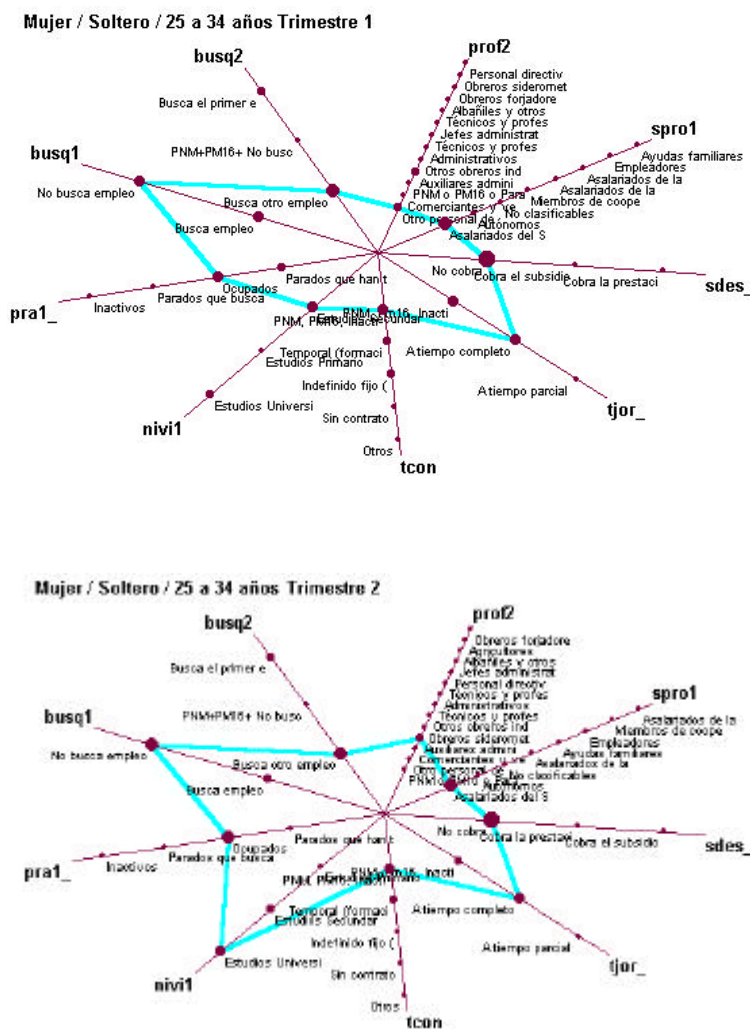
Análisis de la Evolución de un Objeto Simbólico

La representación en forma de estrella puede ser utilizada también para analizar la evolución de un objeto simbólico. Al visualizar diferentes versiones de un mismo objeto pueden identificarse valores estables o valores que muestren una variación grande de una versión a otra.

Ejemplo

Puesto que la encuesta P.R.A. es una encuesta panel que se toma cada trimestre con una rotación de 1/8, vamos a hacer un estudio de la evolución del mismo objeto simbólico en dos trimestres consecutivos.

Se han creado objetos simbólicos correspondientes al cruce de las variables sexo x estado civil x edad. El objeto elegido en los dos trimestres para el estudio es “ Mujer / Soltero / 25 a 34 años” .



Por la forma de la línea que une los ejes, se puede observar que ha habido una evolución en el nivel de estudios (niv1) del colectivo. En el primer trimestre la mayoría poseían estudios secundarios mientras que en el segundo periodo la mayoría tiene estudios universitarios. Otro cambio sucede en la profesión (prof2) donde en el primer trimestre el grupo era mayoritariamente “Otro personal de Servicios” y en el segundo el mayor porcentaje se lo lleva “Auxiliares Administrativos”.

Por el contrario no existen cambios significativos en las variables relación con la actividad, búsqueda de trabajo, situación profesional, tipo de jornada laboral, tipo de contrato,...

Ventajas del uso de Objetos Simbólicos

La visualización de un objeto como estrella permite representar datos complejos con diferentes niveles de detalle.

Esta representación permite la visualización de un objeto cada vez, o varios superpuestos. Esta es la mayor diferencia con respecto a métodos que representan nubes de puntos y los que buscan interpretar interacciones entre variables.

Estadística Básica de Objetos Simbólicos

La estadística básica de Objetos Simbólicos va a consistir en un conjunto de gráficos y medidas resumen que van a depender de que variables formen esos objetos.

Si las variables son multinomiales, se pueden construir gráficos de frecuencias de barras y de sectores para cada una de ellas.

Si las variables son intervalo, se pueden construir gráficos de frecuencias con medidas de tendencia central y dispersión. Además, con estas mismas variables se pueden representar biplots.

Si las variables son probabilísticas, se pueden construir gráficos de capacidades con cada una de ellas.

Frecuencias de Variables Multinomiales

Diagrama de Barras

Son gráficos que representan la distribución que toma una variable multinomial en un conjunto de objetos simbólicos.

Tenemos la variable Clase_ID con las modalidades: pequeño (picc), mediano (medi) y grande (gran). Medimos la frecuencia de esas modalidades en todos los objetos simbólicos de un array dado. En el gráfico nos encontramos que la modalidad que más frecuentemente aparece en todo los objetos simbólicos es “grande”.

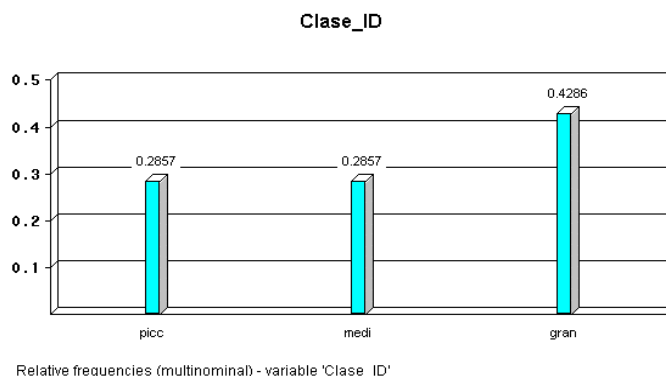
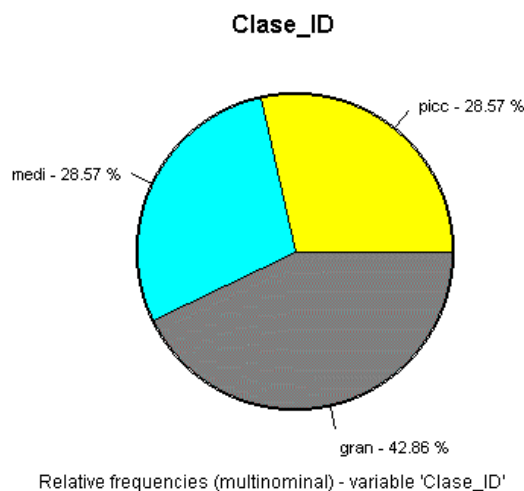


Diagrama de Sectores

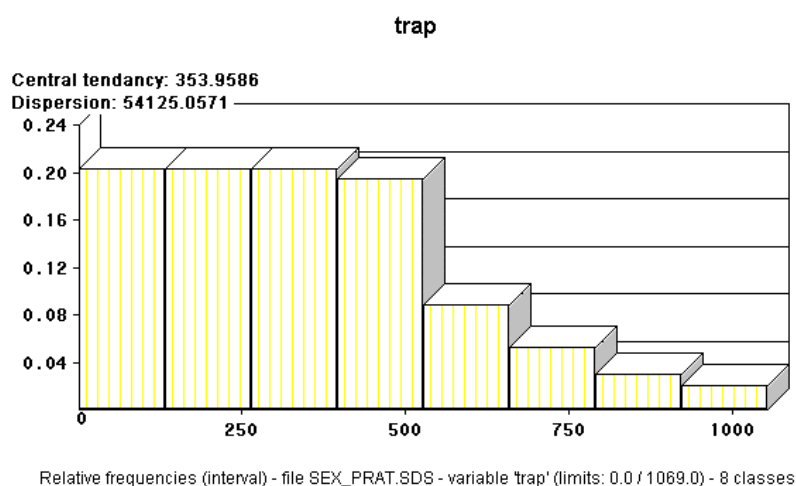
Para el mismo tipo de variable, también puede representarse su distribución mediante un diagrama de sectores.



Frecuencias de Variables Intervalo

Se puede estudiar la distribución de una variable construyendo un histograma "simbólico" dónde los valores de la variable son intervalos. El número de intervalos es elegido por el usuario y van desde el mínimo al máximo valor de la variable elegida.

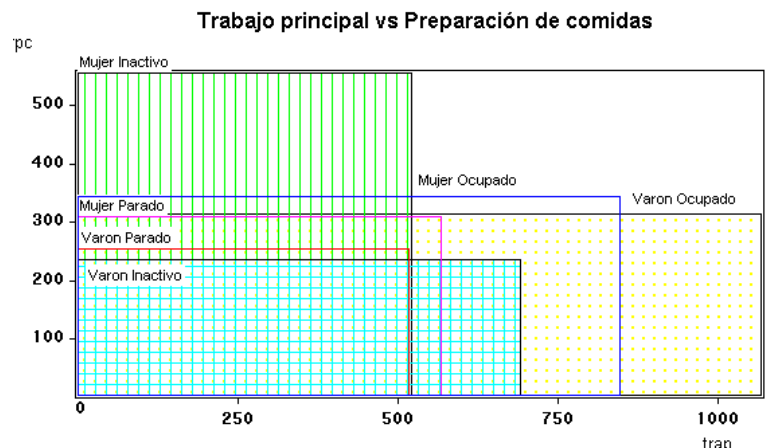
En el gráfico se incluyen los valores de tendencia central y de dispersión de la variable. En este ejemplo se ha elegido la variable "Minutos dedicados al Trabajo Principal al día" de la Encuesta de Presupuestos de Tiempo y se han escogido 8 clases para representarla.



Biplot

Es un gráfico que representa dos variables intervalo. Cada objeto simbólico está representado por un rectángulo en el plano definido por esas dos variables.

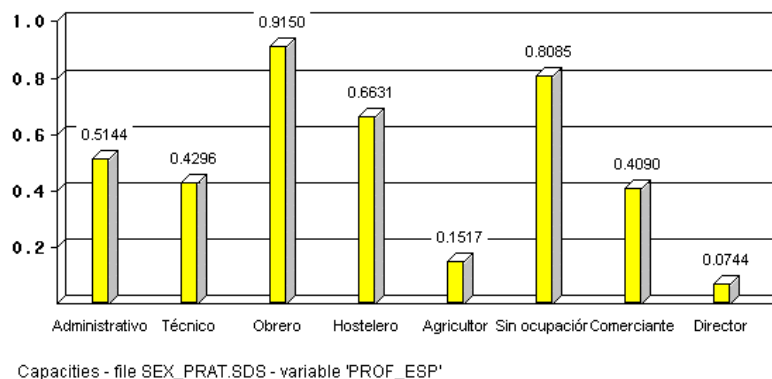
A continuación se han representado las variables de la Encuesta de Presupuestos de Tiempo: Trabajo principal y Preparación de comidas en minutos. Los 6 objetos simbólicos (rectángulos) corresponden al cruce de las variables Sexo x Relación con la Actividad.



Del gráfico puede observarse que el colectivo que más tiempo dedica a la Preparación de comidas es la "Mujer Inactiva" y el que menos "Varón Inactivo". Por el contrario, el que más tiempo dedica al Trabajo principal es "Varón Ocupado" y el que menos "Varón Parado".

Capacidades para Variables Probabilísticas

Este gráfico representa la capacidad de las categorías de una variable para llegar a 1 en un conjunto de objetos simbólicos. En este caso, se ha representado la variable "Profesión" con 8 modalidades de todos los objetos simbólicos generados a partir de la encuesta de Presupuestos de Tiempo.



Conclusiones

Tradicionalmente en la Estadística Oficial se realizan análisis de ficheros planos dónde las unidades estadísticas son individuos. Lo que se ha propuesto en este cuaderno son unas nuevas unidades estadísticas resumen llamadas objetos simbólicos, que contienen más información que las anteriores.

Estas nuevas unidades estadísticas conllevan las siguientes ventajas en la Estadística Oficial:

- a. En los Institutos de Estadística de distintos países se utilizan conceptos casi equivalentes (como son los datos de desempleo o los accidentes de coche), pero estos conceptos pueden estar descritos por diferentes variables y nomenclaturas oficiales $\Rightarrow$ Los Objetos Simbólicos proporcionan una estructura para la *descripción, unificación y análisis* de estos conceptos.
- b. La difusión de resultados es una de las tareas principales de la Estadística Oficial $\Rightarrow$ Los Objetos Simbólicos proporcionan una manera eficiente de *presentación y visualización* de datos con estructura compleja.
- c. Facilidad para transformar los datos en *estructuras de array*.
- d. Los Objetos Simbólicos proporcionan una buena manera de representación de *datos agregados*.

Otros usos de los Objetos Simbólicos

Además de los usos expuestos anteriormente, se están estudiando nuevas vías de aplicación de los objetos simbólicos tales como:

Protección de Datos

En relación con los objetos de tipo intervalo, otra aplicación sería la protección de datos mediante intervalos de garantía. De esta manera, los datos sensibles de una base de datos tendrían asegurada la privacidad si en lugar de dar ese dato como respuesta a una consulta se diera un objeto simbólico de tipo intervalo.

Ejemplo:

Consulta: ¿Años en la empresa y Salario del Sr. López?

Como el resultado de la consulta se considera confidencial, se puede devolver un intervalo en lugar del salario exacto.

Sr. López = [Años = 3] $\wedge$ Salario = [3.5:4.2]

Futuro del Proyecto

La primera etapa del Proyecto SODAS se ha cerrado con el resultado del software SODAS versión 1.04. En este software están implementados los módulos nombrados en el cuaderno, así como varios más que se explicarán en sucesivos trabajos.

Una nueva etapa del proyecto está en marcha, el SODAS II, que mejorará lo alcanzado hasta ahora e incluirá nuevos módulos para el manejo y el análisis de Objetos Simbólicos.

Bibliografía

[1] BERTIER P., BOUROCHE J.M.

Analyse des données multidimensionnelles. ISBN 2 13 0373380. Chapter XI, pages 175-192. Presses Universitaires de France (1981).

[2] BOCK H. H., DIDAY E.,

Analysis of symbolic data. Springer-Verlag (1999)

[3] CHOUAKRIA A.

Mémoire de DEA. Extension de méthodes de réduction de dimension à des données symboliques. Université Paris IX Dauphine- INRIA Rocquencourt (1994).

[4] DIDAY E.

Analyse des données et classification automatique numérique et symbolique. Seminario Internacional de Estadística en Euskadi. Volumen 27. EUSTAT, Euskal Estatistika Erakundea/ Instituto Vasco de Estadística (1992).

[5] DIDAY E.

Extracting Information from very Extensive Data Sets by Symbolic Data Analysis. University Paris 9 Dauphine.

[6] DIDAY E.

From data to knowledge, boolean, probabilist and belief objects for symbolic data analysis. Une introduction a l'analyse des données symboliques. INRIA- Rocquencourt. Domaine de Voluceau. Le Chesnay Cedex. Tutorial at IFCS'93.

[7] DIDAY E.

Users' taking account report.

[8] EUSTAT.

Metodología de la Encuesta de la Población en Relación con la Actividad en el Mercado de Trabajo. EUSTAT, Euskal Estatistika Erakundea / Instituto Vasco de Estadística (1984,1992 y 1997).

[9] HEBRAIL G., LECHEVALLIER Y., STEPHAN V.

SODAS-RDBMS Interface.

[10] IZTUETA A., CALVO P.

Utilities and Applications of Symbolic Data Analysis to Official Statistics.
Congreso KESDA'98, Luxemburgo.

[11] IZTUETA A., CALVO P., LAAKSONEN S., DIDAY E.

Uses of Symbolic Objects in Official Statistics. Congreso ISI'99, Helsinki.

[12] LEBART L., MORINEAU A., PIRON M.

Statistique exploratoire multidimensionnelle. Chapter 3, Section 5, pages 302-318.
Dunod, Paris (1995).

[13] PERINEL E.

Analyse numerique/symbolique des tactiques de pêche artisanale au Senegal.
D.E.A. de Mathématiques Appliquées aux Sciences Economiques. Université Paris-IX Dauphine et ORSTOM (1991-92).

[14] REYNER A.

Analyse Symbolique de Scénarios d'Accidents. Université Paris XI Dauphine-INRETS
(1991-92).

[15] SODAS Project.

Informes Internos. Contract N°20821 DG34/D-3/300536.

[16] SODAS Project.

SUM (Software User Manual).

[17] STÉPHAN V.

Creation d'assertions à partir de requêtes ORACLE. (1994).

[18] STÉPHAN V.

Extracting Symbolic Objects from Relational Databases.

[19] STÉPHAN V., HEBRAIL G., LECHEVALLIER Y.

Building Symbolic Objects from Relational Databases.

[20] STÉPHAN V., HEBRAIL G., LECHEVALLIER Y.

Improving Symbolic Descriptions of sets of Individuals: the reduction of assertions.