APPLICATIONS OF SYMBOLIC OBJECTS IN OFFICIAL STATISTICS

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

APPLICATIONS OF SYMBOLIC OBJECTS IN OFFICIAL STATISTICS

Patricia CALVO GARRIDO

Statistical Assistant Technician at EUSTAT

SUMMARY

This study mainly deals with the definition, creation and visualization of Symbolic Objects from surveys stored in Relational Databases. In addition, a chapter is dedicated to the basic statistics of these Objects, leaving more detailed analyses for further technical notebooks. At each step, the advantages of this new kind of analysis are presented in comparison with a classical analysis.

The use of symbolic objects in Official Statistics provides the capacity to describe data with a complex structure. Information on the structure will be extracted from the database and will be described by means of symbolic objects. These objects may describe individuals or kinds of individuals.

The structure of the notebook is as follows:

In the *Introduction* the SODAS project is presented, on which this notebook is based, the basic notions of the definition of a symbolic object and the origin of this new notation.

Chapter Two describes how knowledge is extracted from a database by means of queries and how the result of these queries is transformed into symbolic objects.

In *Chapter Three* we look at the way to visualize symbolic objects through tables, graphs or the exclusive language of symbolic objects.

Chapter four focuses on the descriptive statistics of symbolic objects

Finally, the conclusions of this study are presented together with the future of this new theory in official statistics.

KEY WORDS:

Symbolic Object, assertion, Boolean object, probabilistic object, relational database, query.

Index

Chapter
1

Introduction

Introduction and objectives

The aim of this paper is an initiation to Symbolic Data Analysis in Statistical Offices. In this case, we have applied this new analysis to the study of some EUSTAT surveys to know the advantages with regard to a classical data analysis.

Classical methods of statistical data analysis were designed for a relatively simple situation. First, data were obtained for single individuals using interviews, experiments, archives, etc. Second, the variables were well defined and third, these variables had only a value or category in each individual.

Sometimes, the real world is too complex to be described by these relatively simple models. In order to deal with these cases, we will introduce the symbolic object concept and define various types of symbolic data.

A Symbolic Object is a way of representing complex data that arise when analysing huge data sets. In Statistical Institutes one of the most important tasks is to summarize them in shorter sets with new statistical units, losing as little information as possible. The new statistical units will be the called Symbolic Objects and will extend the Standard Data Analysis to the corresponding Symbolic Data Analysis.

Project Description

The use of Symbolic Objects proposed by E. Diday, has reached its maximum development in the framework of the European project SODAS.

“SODAS: Symbolic Official Data Analysis System” is project no. 20281 of the European Commission, General Directory III, Industrial rtd, EUROSTAT, DOSES program.

In this project, several members belonging to Universities, Enterprises, Official Statistics and Research Centres of the European Union are involved, one which is the EUSTAT, an official statistical office.

The aim of this project is to facilitate the use of Symbolic Data Analysis in statistical offices and companies, and consequently demonstrate that these techniques meet several user needs:

- Analysis of data with complex structures.
- Better explanations of statistical results.
- Concepts and metadata representation, manipulation and analysis.

- Exchange of data between official statistics community members.

This will be achieved by:

- Developing the software for prototyping and evaluating statistical processing of symbolic data.
- Building a concept oriented system for official statistics.

The software will include:

- Generic tools for storing, querying, and updating symbolic objects.
- Tools for acquiring symbolic objects from large databases.
- A collection of data analysis methods dedicated to symbolic objects: univariate descriptive methods, clustering, decision-tree construction, discrimination, and factorial analysis.
- Facilities to transform symbolic objects into “standard” objects and then perform standard data analysis methods on them.

Ergonomic tools for presenting to the user the results of the methods.

Need for Symbolic Objects

The following are examples that illustrate the need to use symbolic objects:

For example, for individuals, the variable Y = "Minutes dedicated to practising sport per day" is a variable that allows a non-unitary response, as it varies from day to day. For an individual k, this variable may be expressed in a non-classical manner:

Y(k) = [20,60] o

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

Y(k) = {Null Participation (0.1), Low Particip. (0.5), Average Particip. (0.3), High Particip. (0.1)}.

For groups of individuals, if k denotes the region ‘Álava’, the variable Y = "Relation with the Activity" may be specified by:

Y(k) = {Employed (0.47), Unemployed (0.11), Inactive (0.42)},

which means that 47% of the individuals in Álava are employed, 11% unemployed, etc..

Data Tables

The creation of Symbolic Objects is based on tables from a Relational Database where,

- There are various related tables at different levels.
- Data and the metadata appear separately, so information is not repeated.
- It is possible to have one data table and several metadata tables in different languages.

The generated symbolic objects will be also stored in tables. These tables, with symbolic objects in rows and variables in columns, will be the beginning to different algorithms of Symbolic Data Analysis. Each cell of these tables may contain different types of data, such as:

a) A single quantitative value: age (w) = 23;

b) A single categorical value: sex (w) = woman;

c) Multivalued: marital status (w) = {single, married}.

d) Interval: age (w) = [20, 25];

e) Multivalued with weights: age (w) = [20 (0.65), 25 (0.35)];

Where *age, sex* and *marital status* are variables and *w* units.

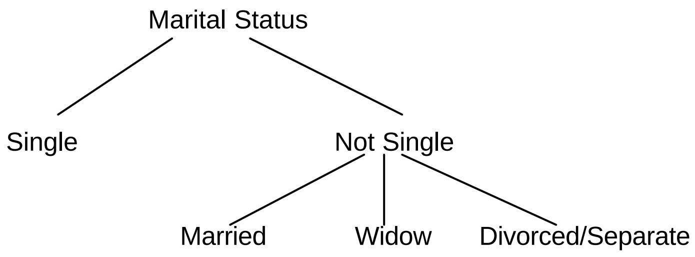
In short, the values that individuals take in the variables can be *non-atomic* (a group of values, an interval or a probability distribution).

Table 1. Table of Symbolic Objects

	Sex	**Age**	**Relation to Activity**
OS 1	{woman (0.33), man(0.67)}	{[25:57]}	{employed(0.67), inactive(0.33)}
OS 2	{woman (0.25), man(0.75)}	{[15:42]}	{employed (0.25), unemployed (0.25), inactive(0.50)}
OS n	{woman (0.5), man(0.5)}	{[27:29]}	{employed (1)}

The variables that describe symbolic objects may be:

1. Variables with Taxonomic Domain: Offer the possibility of defining a taxonomy within the values taken by a variable. This taxonomy represents a priori knowledge on the data.



The same as a symbolic object:

marital_status= {single, not single = {married, widow, divorced/separate}};

2. Mother-Daughter Variables (or Hierarchically Dependant): Offer the possibility of defining variables which are not applicable to all the individuals, but only to the individuals verifying some properties.

   IF *Relation to Activity = unemployed* THEN *Type of Contract* is N.A. (not applicable);

3. Variables with Logical Dependencies (or rules): Offer the possibility of defining a priori knowledge on the data in the form of a restriction of the possible combinations of values for the different variables.

   IF *age>65* THEN *Professional Situation = Retired*

Depending on the type of data that compound the symbolic objects, these can be of different types:

- Boolean Objects: if the associated cells are only of type a), b), c) and d), above described.

- Modal Objects: if at least one cell of the corresponding row contains modes.

In summary, the scheme process for obtaining Symbolic Objects is:

Queries to a Database → Building Symbolic Objects → Symbolic Tables → Symbolic Data Analysis.

Formal Definition of Symbolic Object

Let us define a ***Symbolic Object*** as *"a description expressed with a set of events (or properties) induced by the values taken by the variables"* (see [4]).

A variable y is a mapping $\Omega \rightarrow V$ where Ω is the set of "elementary objects" and O is the observation set where the variable takes its values ($V \subset O$).

Types of Symbolic Objects

- Elementary events: $\mathbf{e_i = [y_i=V_i]}$
It is a mapping e_{y_i}vi: $\Omega \rightarrow \{$true, false$\}$ such that $e_{y_i}V_i(w)$=true if and only if $y_i(w) \in V_i$.

Ex. [length = 0.52] or [colour $\in$ {red, blue}]

- Assertion: $\mathbf{a = [y'_1=V_1]}$ Ù...Ù $\mathbf{[y'_q=V_q]}$ where $V_i \subset O'_i$ is defined by the mapping $a_y V$: $\Omega \rightarrow$ {true, false} such that $a_y V(w)$=true if and only if for all i=1,...,q $y_i(w) \in V_i$.
It is a conjunction of events that have to be true simultaneously for the same elementary object $w \in \Omega$.

Ex. [length = 0.52] $\wedge$ [colour $\in$ {red, blue}] $\wedge$ [shape = rectangular]

- Horda Object: $\mathbf{h = [y'_1(u_1)=V_1]}$ Ù...Ù $\mathbf{[y'_p(u_p)=V_p]}$ is defined by the function $h_y V$: $\Omega^q \rightarrow$ {true, false} such that $\forall$ W = $(w'_1,\ldots,w'_q) \in \Omega^q$, $h_y V(W)$ = true if and only if $\forall$ i $y'_i(w'_i) \in V_i$.

- Synthesis Object: $\mathbf{s = h_1}$Ù... Ù$\mathbf{h_k}$**,** is the conjunction of k horda objects defined respectively on each of the groups $H_1,\ldots,H_k$ con $h_i \in H_i$.

Modal Symbolic Objects

Previously, we defined modal objects as having at least one cell with weights. Now, we are going to classify these objects according to whether the weights (or modes) affect the whole object or only its values.

External Modal Objects: the modes affect the event globally.
$\mathbf{a_x}$ **= î** $\mathbf{M_i}$ $\mathbf{[y_i=V_i]}$ where x refers to the semantic (possibilistic, probabilistic,...).

Ex.: *often* [Age = [16,24]] (possibilistic)

Internal Modal Objects: the modes affect the values taken by the variables.
$\mathbf{a_x}$ **= î** $\mathbf{[y_i= M_i\ V_i]}$ where x refers to the semantic.

Ex.: [Marital_status = not single *(0.7)*, single *(0.3)*] (probabilistic)

Definition of an Object in Intension and in Extension

Definition of an object in Intension*:* The object is described by the properties that characterize it.

Definition of an object in Extension: The extension of a symbolic object is the set of elementary objects of Ω that satisfy it. We denote $|s|_\Omega$ or Ext(s).
In the Boolean case, Ext(s) ={$w \in \Omega$ / a(w) = true}
In the modal case, given a threshold α, Ext(s) = {$w \in \Omega$ / a(w) $\geq \alpha$}

In this work, we will use internal modal elementary events and assertions defined in Intension.

Example

We have the following table with 4 individuals (in rows) and 3 variables (in columns):

	y_1: Sex	y_2: Age	y_3: Education
Ind1	Woman	25	university
Ind2	Woman	60	primary
Ind3	Man	38	secondary
Ind4	Man	54	secondary

The set of "elementary objects" is Ω ={Ind1, Ind2, Ind3, Ind4}.
The observation set of the variable y_1 is O_1= {woman, man} and in the same way for y_2 and y_3.

Then, Ind1 may be described by the following Boolean assertion:
Ind1 = [Sex = woman] $\wedge$ [Age = 25] $\wedge$ [Education = university]

To clarify the terms of definition in intension and extension, another symbolic object could be:
a = [Sex = man] $\wedge$ [Education = secondary].
This description is a definition in intension.
Its definition in extension would be:
Ext(a) = {Ind3, Ind4}, since individuals 3 and 4 fulfil 'a'.

First and Second Order Symbolic Objects

First Order

Symbolic Objects are first order types when the data refer to single individuals.

Let be E = Ω ={1,...n} a universe of individuals (elementary objects).

For example, the variable Y = "Age" for each pupil k at a school:

Y(k) = {11} o

Y(k) = [4, 13]

Second Order

Symbolic Objects are second order types when the data refer to more or less homogeneous classes of individuals. As not all individuals of the same class take the same value in each variable, there might be several categories which apply simultaneously to the class, eventually with specified percentages.

Let be E = {C_1, C_2,...} a system of classes $C_i \subseteq \Omega$ (aggregated objects).

Now k denotes a class of individuals such as a specific year at the school in the above example and the variable Y = "Age" may be specified by:

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

which means that 20% of the individuals of this year are aged 10, 60% are 11 years old, etc..

Higher order objects can be defined in an analogous way by successive aggregation steps (aggregate 2 different years of the same level and compare them on the basis of age).

Symbolic Objects as an Imprecise Measure

We can find studies that cannot be based on unique experimental or interviewing results, but take into account some inaccuracy. It is here where other types of symbolic objects based on imprecise results appear. This includes probabilistic o possibilistic data, fuzzy data, or interval data.

Intervals may result from two sources: from observations or directly from expert knowledge. In the case of data resultant from observations or measures there are, on the one hand, intervals due to imprecise knowledge: the result ξ_{ij} of an observation or measure is an interval $[a_{ij}-\delta, a_{ij}+\delta]$ where a_{ij} is the observed value and δ characterizes the imprecision of the measure instrument. On the other hand, there are intervals due to variability: let $a_{ij}^{1},\ldots, a_{ij}^{k}$ be observations of variable j for object i. The result of summarizing these k observations is the interval $x_{ij} = \left\lfloor \underline{x_{ij}}, \overline{x_{ij}} \right\rceil$ where $\underline{x_{ij}}, \overline{x_{ij}}$ are the minimum and maximum observed values, respectively.

Example

We consider the estimations of a measure by means of confidence intervals in groups defined by territory, relation to labour activity and branch of economic activity. As we do not know the exact measure, we will define a symbolic object that includes the interval measure.

Y = [terr = Alava] $\wedge$ [pra1 = Employed] $\wedge$ [ract2 = Agriculture, cattle and fishing] $\wedge$ [estimation = [32.75, 40.8]].

i.e., the estimation for this collective will be between 32.75 and 40.8.

An expert is not 100% sure about his affirmations and in this case he expresses doubts, beliefs, etc. Usually, intervals are used to describe expert knowledge including uncertainty.

Source of Symbolic Objects

Symbolic Objects result from many methods:

1) From Relational Database Queries.

2) From Data Analysis of standard tables to build groups (factorial analysis, clustering, …).

3) From expert knowledge.

4) From time series.

5) From confidential data (to hide initial data by means of less accuracy).

In our case, using some surveys we will generate symbolic objects by means of queries to the database.

In the following chapters different types of symbolic objects will be developed, insisting on the advantages with respect to traditional methods of data analysis.

Chapter
2

Obtaining Symbolic Objects from Database Queries

The most direct way to obtain symbolic objects is by means of database queries. These queries automatically extract groups of individuals with common characteristics, for example families, regions, etc. It is really a generalization process of a group of classical data stored in a database bearing in mind relations between different tables. Once the symbolic objects have been created, a process of specialization may be applied to them in order to reduce over-generalization or to join several together by applying a joint operator.

Building Assertions

A relational database follows a table structure where each tuple represents an individual.

A way to obtain and describe information stored in a relational database is building symbolic objects. These objects are created by aggregating individuals in classes and describing the properties of these classes.

In the selection process for the population we take into account data of several related tables with additional knowledge such as taxonomies, mother-daughter variables. The steps of the process are:

- SQL query in DB2SO of SODAS that specifies which relevant data need to be processed and which attributes need to be returned.

  The general format of a SQL query is:

  ```
  SELECT id, group attribute, rest of variables, [sampling weight]
  FROM table
  WHERE restrictions;
  ```

  These queries consist of a unique id for each individual, an attribute that groups the individuals (*group attribute*), other variables that show the composition of the group, and optionally a sampling weight. The composition of the group can be given in percentage or in figures.

  The result of a query, i.e. a group of tuples, is considered as the population under study. If the size of the population is great, random sampling may be carried out.

- Description of each group by an assertion, for future analysis of these groups.

Simple Group Attribute

The aggregation is made by a unique variable. We will obtain an equal number of symbolic objects as the number of categories belonging to the variable.

Composed Group Attribute

The group attribute is composed of two or more categorical variables. We will obtain as many symbolic objects as the product of all the variable categories.

Queries with Restrictions

We impose a filter on the data to be returned by the query. The restriction can affect the group attribute or any other variable that describes the group.

Examples

<u>Simple Group Attribute</u>:

SELECT id, ***marital_status***, sex, level of education, relation to activity, age, …
FROM table

With this query we will obtain 4 symbolic objects that describe the marital status of the population: “Single”, “Married”, “Widow”, “Divorced/Separate”, and that will be described by the rest of variables sex, age, …

```
os "Married"(1684) =
   [sexo = {"Man"(0.50822), "Woman"(0.49178)}]
  ^[nivi1 = {"University studies"(0.10457), "Secondary studies"(0.240049),
"Primary studies or less"(0.655381)}]
  ^[pra1 = {"Employed"(0.497875), "Unemployed having worked"(0.0701014),
"Inactive"(0.428778), "Unemployed seeking employment"(0.00324637)}]
  ^[eden = {"16 to 24 years old"(0.00694884), "65 and more years
old"(0.191936), "45 to 54 years old"(0.249274), "55 to 64 years
old"(0.182608), "25 to 34 years old"(0.106163), "35 to 44 years
old"(0.263071)}]

os "Single"(1001) =
   [sexo = {"Man"(0.534476), "Woman"(0.465524)}]
  ^[nivi1 = {"University studies"(0.230614), "Secondary studies"(0.518448),
"Primary studies or less"(0.250937)}]
  ^[pra1 = {"Employed"(0.475182), "Unemployed having worked"(0.118261),
"Inactive"(0.335337), "Unemployed seeking employment"(0.0712199)}]
  ^[eden = {"16 to 24 years old"(0.442392), "65 and more years old"(0.031812),
"45 to 54 years old"(0.0404278), "25 to 34 years old"(0.375095), "55 to 64
years old"(0.0367008), "35 to 44 years old"(0.0735729)}]
```

<u>Composed Group Attribute</u>:

SELECT id, ***marital_status & sex***, level of education, relation to activity, age,…
FROM table

In this case, we will obtain 8 (4*2) symbolic objects combination of categories of marital status with those of sex: “Single Man”, “Single Woman”, “Married Man”, “Married Woman”, “Widow”, “Widower”, “Divorced/Separate Man”, “Divorced/Separate Woman”.

```
os "Married / Man"(844) =
   [nivi1 = {"Primary studies or less"(0.622043), "Secondary
studies"(0.25979), "University studies"(0.118167)}]
  ^[pra1 = {"Employed"(0.677499), "Unemployed seeking
employment"(0.00130055), "Inactive"(0.287075), "Unemployed having
worked"(0.034125)}]
  ^[eden = {"16 to 24 years old"(0.00517072), "55 to 64 years old"(0.188175),
"25 to 34 years old"(0.0798497), "65 and more years old"(0.205099), "35 to 44
years old"(0.264568), "45 to 54 years old"(0.257138)}]

os "Married / Woman"(840) =
   [nivi1 = {"Primary studies or less"(0.689833), "Secondary
studies"(0.219648), "University studies"(0.0905192)}]
  ^[pra1 = {"Employed"(0.312245), "Unemployed seeking
employment"(0.00525723), "Inactive"(0.575217), "Unemployed having
worked"(0.10728)}]
  ^[eden = {"16 to 24 years old"(0.00878641), "55 to 64 years old"(0.176855),
"25 to 34 years old"(0.133355), "65 and more years old"(0.178333), "45 to 54
years old"(0.241147), "35 to 44 years old"(0.261524)}]
```

Query with Restrictions:

SELECT id, ***marital_status***, sex, level of education, relation to activity,…
FROM table
WHERE ***25 < age < 35***

We will obtain 4 symbolic objects describing each marital status as in the first example, but now individuals are between 25 and 35 years old.

```
os "Married"(150) =
   [nivi1 = {"Secondary studies"(0.442944), "University studies"(0.13868),
"Primary studies or less"(0.418376)}]
  ^[pra1 = {"Unemployed seeking employment"(0.0059544), "Inactive"(0.148825),
"Employed"(0.665266), "Unemployed having worked"(0.179955)}]

os "Single"(329) =
   [nivi1 = {"Secondary studies"(0.517487), "University studies"(0.292927),
"Primary studies or less"(0.189587)}]
  ^[pra1 = {"Unemployed seeking employment"(0.0723981),
"Inactive"(0.0509509), "Employed"(0.684991), "Unemployed having
worked"(0.19166)}]
```

It can be seen that with the restriction on the age, the number in each object decreases. Compare the numbers in brackets from the first example objects with these.

Building Mother-Daughter Variables or Hierarchical Dependencies

As we saw in chapter 1, mother-daughter variables define variables that are not applicable to all individuals, but only to the individuals verifying some properties.

Some of the variables that we treat in the surveys depend on the answer of previous variables.

For example, the variable busq2 (seeking the first employment or other) in the P.R.A survey, is only valid for individuals that have declared that they are seeking employment. It has no sense to apply this variable to an individual if another variable (busq1) shows that the individual is not looking for employment. The rule in this case will be, IF busq1 = “Not seeking employment” THEN busq2 = N.A (Not applicable)

Building Mother-Daughter variables transforms the individuals file. The daughter variable (in this case busq2) now depends on the values taken by the mother variable. If the mother variable takes the value where the daughter is applicable (*seeking employment*), then the daughter variable will take one of its possible values (*first employment, other employment,...*). However, if the mother takes values where the daughter is *not applicable* (*not seeking employment*), the latter will take the value N.A.

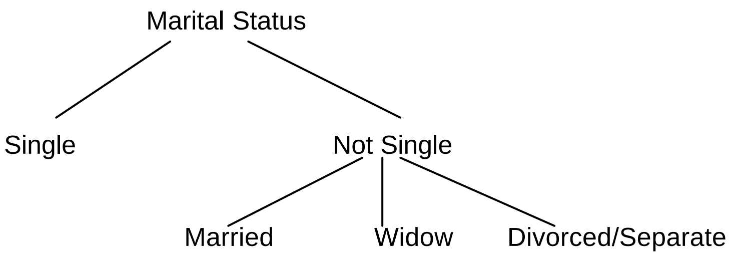
The new rule appears in the *assertions file* and it is made positive, i.e., where it is applicable instead of where it is not applicable.

```
busq2 is applicable if busq1 in {“Seeking employment”}
```

Addition of Taxonomies in the variable domain

The taxonomic variables allow us to define an order in their values. The definition of the hierarchy of values needs a priori knowledge.

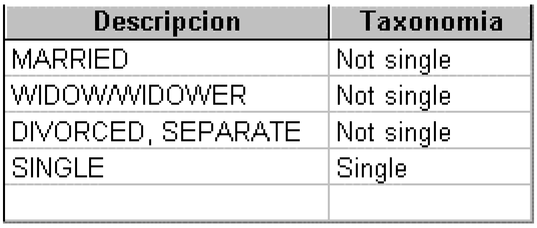
For example, we can define an order in the values of the variable *marital status:*



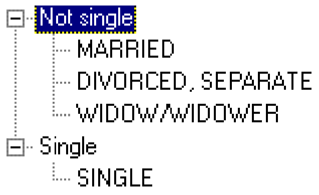
Before the SQL query, we have to build a table in the database defining the taxonomy.

For the previous example about marital status, we show the database table and the visualization in the DB2SO module once the taxonomy has been created.

Table in the Database **Visualization of the taxonomy**



Descripcion	Taxonomia
MARRIED	Not single
WIDOW/WIDOWER	Not single
DIVORCED, SEPARATE	Not single
SINGLE	Single



Not single
MARRIED
DIVORCED, SEPARATE
WIDOW/WIDOWER
Single
SINGLE

In the assertions file, the description of the variable including the taxonomy can be seen.

```
variable eciv2
  nominale {"SINGLE", "MARRIED", "DIVORCED, SEPARATE",  "WIDOW/WIDOWER",
"Single", "Not single","root_eciv2"}
  multiple,mode=probabilist;
  taxonomy {
    "Single" = {"SINGLE"},
    "Not single" = {"MARRIED", "DIVORCED, SEPARATE", "WIDOW/WIDOWER"},

    "root_eciv2" = {"Single", "Not single"}
  }
```

Refinement of Assertions

The assertions described above, considering all individuals, even atypical ones, derive in over-generalization. For that reason, we now propose a new approach based on volume where the final description rejects atypical individuals.

To obtain more homogeneous assertions, we refine them. This procedure decreases the volume of assertions removing some individuals in each group with a minimum covering power. In addition, the number of overlaps between assertions decreases.

The assertions become more specific, easier to interpret and give a better description of the characteristics in terms of homogeneity. The quality of the assertions produced is of great importance as they are used as input for symbolic data analysis methods.

To perform the reduction step, we adapt a volume criterion that measures a generality index on the union of individuals:

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

This volume criterion cannot be applied with both numeric and nominal features. To overcome problems of scale between numeric and nominal features, we transform numeric features into ordinals. This coding allows us to have a homogeneous criterion among all kinds of features without privileging one kind in particular.

We code a numeric domain feature by searching for a uniform distribution over each interval of points. So, in the numeric case d_i is a set of “uniform” intervals, which generalize individual descriptions. We also weight generalization over taxonomic features to privilege the simplicity of description.

We fix a threshold α, which is the minimum covering power of the assertion. The covering power of an assertion a_C is computed with a, and the function of membership corresponding to a_C is:

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

where

$$a(w) = \begin{cases} 1 \quad if \quad w \quad belongs \quad to \quad description \quad a \\ \qquad\qquad 0 \quad otherwise \end{cases}$$

Each step of the algorithm removes from the description the attribute value which maximizes the reduction of hyper-cube a under α covering constraint.

Example

We have applied a reduction with a α=85% to the following example:

```
os "Woman 25 to 34 years old University studies Employed"(39) =
   [eciv2 = {"Married"(0.24562), "Single"(0.75438)}]
  ^[prof2 = {"Superior technicians and professionals"(0.179049), "Merchants
and salespersons"(0.0999308), "Other personnel of Services"(0.0241),
"Administrative managers"(0.0459999), "Administrative staff"(0.0320081),
"Administrative auxiliaries"(0.166976), "Managers"(0.0239909), "Medium
technicians and professionals"(0.427945)}]
  ^[spro1 = {"Hired in private sector"(0.468736), "Hired in public
administration"(0.427499), "Self-employed"(0.0258654), "Hired in public
company"(0.0778989)}]
  ^[tjor = {"Full-time (3 or more h/day)"(1)}]
  ^[tcon = {"Others"(0.0494191), "NSP+YT16+Unemployed+Inactive+Not hired,
Cooperative member"(0.0258654), "Indefinite fixed (permanent or
discontinuous)"(0.401042), "Temporary (training, seasonal, occasional or
others)"(0.523673)}]
```

After reduction, the assertion has 5 less individuals considered atypical:

```
os0.85 "Woman 25 to 34 years old University studies Employed"(34) =
   [eciv2 = {"Single"(0.772791), "Married"(0.227209)}]
  ^[prof2 = {"Administrative managers"(0.0528245), "Medium technicians and
professionals"(0.466286), "Other personnel of Services"(0.0276755),
"Merchants and salespersons"(0.0850538), "Superior technicians and
professionals"(0.176412), "Administrative auxiliaries"(0.191749)}]
  ^[spro1 = {"Hired in public administration"(0.461722), "Hired in private
sector"(0.538278)}]
  ^[tjor = {"Full-time (3 or more h/day)"(1)}]
  ^[tcon = {"Temporary (training, seasonal, occasional or others)"(0.576216),
"Indefinite fixed (permanent or discontinuous)"(0.423784)}]
```

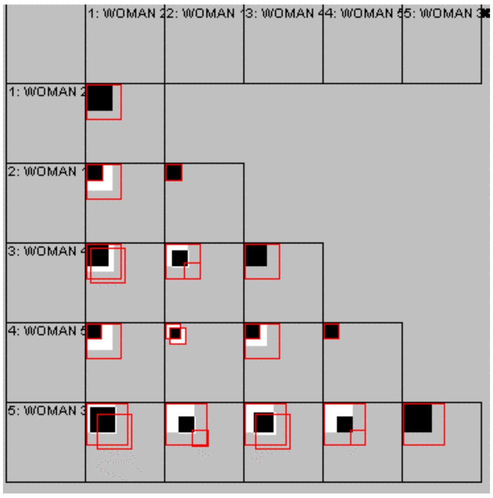


Fig.1: Volume matrix before (blank squares) and after refinement. Black squares show the overlapping between assertions.

Joining Assertions

We consider two different arrays of assertions, but describing the same objects. For instance, one array of assertions describes regions by household variables, and the other one describes the same regions by people employment variables.

Let X_1 and X_2 be two arbitrary symbolic data arrays with individuals corresponding to sets E_1 and E_2 respectively, and with variables $Y_{11},\ldots, Y_{ip}$ y $Y_{21},\ldots Y_{2q}$ respectively. The joining of X_1 and X_2 is denoted by ***join(X*** $_1$***, X*** $_2$***)*** and is a symbolic data array defined as follows:

1. $E = E_1 \cap E_2$ (the set of entities or symbolic objects of the resulting symbolic data array is the intersection of the two sets of entities on which X_1 and X_2 are based).

2. The variables describing *join*(X_1, X_2) are $Y_{11},\ldots, Y_{ip}, Y_{21},\ldots Y_{2q}$ (the concatenation of the variables describing X_1 and X_2).

3. For each $u \in E$ we define *join*(X_1, X_2)(u) := (X_1(u), X_2(u)). Thus the resulting data matrix X = *join*(X_1, X_2) has the format |E| x (p+q).

4. Possible taxonomies defined on some variables of X_1 or X_2 are maintained in *join*(X_1, X_2).

5. Possible mother-daughter variables defined by rules on X_1 or X_2 are maintained in *join*(X_1, X_2).

Entities which show up in X_1 (resp. X_2), but not in X_2 (resp. X_1) are lost in *join*(X_1, X_2).

Uses of Joining Objects in Official Statistics: Data Fusion

We find a new application of joining symbolic objects that consists of joining assertions coming from different surveys. This joining enables us to obtain additional information, data imputation, to obtain conclusions about causes and possible effects,...

Fusion using symbolic objects differs from the traditional data fusion in the way of matching. Instead of joining record by record of common variables, we join by symbolic objects each one describing a group.

The fusion allows us to relate independent surveys to some common items. In the SODAS project framework, this comparison will be between surveys of different countries of the European Union.

Example

An example showing this new use in EUSTAT, is the fusion of the two independent surveys Use of Time (EPT) and Living Conditions (ECV). They have common variables (socio-demographic) and it is probable that there is a relation between them.

The first step is to define the common socio-demographic variables and to create assertions for each survey separately. The group attribute for these assertions will be the concatenation of common variables.

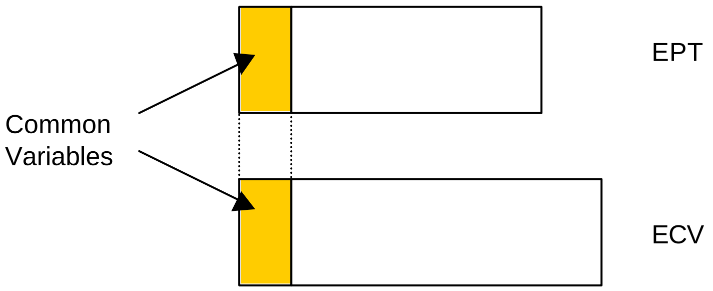


Fig.2: Common part and specific part of two independent surveys.

The common variables chosen for this study were: Sex, Marital status, Age, Relation to Activity and Level of Education.

The second step is to join assertions describing the same group. Then, for the same group we will have the description in the specific variables of each survey.

We consider the following data arrays:

X_1 is a symbolic data array that describes socio-demographic groups by the following variables of Use of Time:

Y_{11}(limp) = Participation in Cleaning

Y_{12} (prpc) = Participation in Preparing Meals

Y_{13} (prac) = Sport Practice

Y_{14} (cuip) = Time used in Personal Care

One of the objects of the array is:

```
os "Woman Married < 35 years Employed Secondary"(54) =
   [limp = {"Null Particip."(0.347273), "Low Particip."(0.188186), "Average
Particip."(0.346782), "High Particip."(0.117759)}]
  ^[prpc = {"Null Particip."(0.0719004), "Low Particip."(0.400066), "Average
Particip."(0.436589), "High Particip."(0.0914451)}]
  ^[prac = {"Null Particip."(0.877218), "Low Particip."(0.122782)}]
  ^[cuip = [0:170]]
```

X_2 is a symbolic data array that describes the same socio-demographic groups by the following variables of Living Conditions:

Y_{21} (jorna) = Length of Working Day

Y_{22} (comt) = Return home to have lunch

Y_{23} (distr) = Distance to Place of Work

Y_{24} (ractp) = Branch of Economic Activity

```
os "Woman Married < 35 years Employed Secondary"(34) =
   [jorna = {"SPLIT SHIFT"(0.394297), "CONTINUOUS"(0.434047), "NOT
APPLICABLE"(0.171656)}]
```

```
  ^[comt = {"RETURN HOME TO LUNCH"(0.637714), "NOT RETURN HOME TO
LUNCH"(0.345755), "NOT APPLICABLE"(0.0165312)}]
  ^[ractp2 = {"PAPER-GRAPHIC ART"(0.0165312), "CONSTRUCTION AND CIVIL
WORKS"(0.0235133), "COMMERCE-HOSTELRY-REPARING-RECOVERY"(0.337456),
"TRANSPORTS AND COMMUNICATION"(0.0317708), "BANK AND INSURANCES"(0.048266),
"NON-COMMERCIAL SERVICES"(0.078488), "PUBLIC ADMINISTRATION-
TEACHING"(0.137167), "VEHICLES AND TRANSPORT MATERIAL"(0.0167604),
"CHEMISTRY"(0.0331782), "COMMERCIAL SERVICES"(0.140833), "RUBBER AND PLASTIC
TRANSFORMATIONS"(0.0199348), "AGRICULTURE-CATTLE-FORESTRY-
FISHING"(0.0201623), "METALLIC CONSTRUCTION"(0.0246369), "ELECTRIC MATERIAL
AND MACHINERY"(0.0497522), "WOOD-FURNITURE"(0.0215493)}]
```

Then, the joint symbolic objects is:

```
os "Woman Married < 35 years Employed Secondary"(88) =
   [limp = {"Null Particip."(0.347273), "Low Particip."(0.188186), "Average
Particip."(0.346782), "High Particip."(0.117759)}]
  ^[prpc = {"Null Particip."(0.0719004), "Low Particip."(0.400066),"Average
Particip."(0.436589), "High Particip."(0.0914451)}]
  ^[prac = {"Null Particip."(0.877218), "Low Particip."(0.122782)}]
  ^[cuip = [0:170]]
  ^[jorna = {"SPLIT SHIFT"(0.394297), "CONTINUOUS"(0.434047), "NOT
APPLICABLE"(0.171656)}]
  ^[comt = {"RETURN HOME TO LUNCH"(0.637714), "NOT RETURN HOME TO
LUNCH"(0.345755), "NOT APPLICABLE"(0.0165312)}]
 ^[ractp2 = {"BANK AND INSURANCES"(0.048266), "NON-COMMERCIAL
SERVICES"(0.078488), "PAPER-GRAPHIC ART"(0.0165312), "COMMERCIAL
SERVICES"(0.140833), "CHEMISTRY"(0.0331782), "VEHICLES AND TRANSPORT
MATERIAL"(0.0167604), "COMMERCE-HOSTELRY-REPARING-RECOVERY"(0.337456), "METALLIC
CONSTRUCTION"(0.0246369), "ELECTRIC MATERIAL AND MACHINERY"(0.0497522),
"TRANSPORTS AND COMMUNICATION"(0.0317708), "WOOD-FURNITURE"(0.0215493),
"AGRICULTURE-CATTLE-FORESTRY-FISHING"(0.0201623), "PUBLIC ADMINISTRATION-
TEACHING"(0.137167), "CONSTRUCTION AND CIVIL WORKS"(0.0235133), "RUBBER AND
PLASTIC TRANSFORMATIONS"(0.0199348)}]
```

Advantages of using Symbolic Objects

Each tuple resulting from a query to the database is converted into a new statistical unit called a symbolic object.

With this new statistical unit, all kind of statistical analyses could be carried out, with the advantage that we can treat groups instead of individuals. Those groups may contain information from several related tables.

Chapter
3

Visualization of Symbolic Objects. Zoom Star.

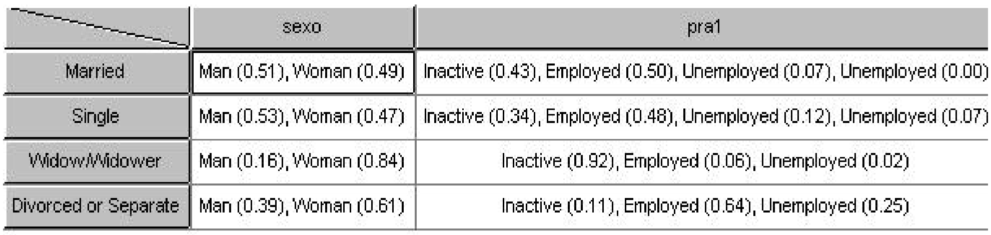
Symbolic Objects may be visualized in three different ways:

- In a symbolic table,
- By star graphs and
- By the specific language of symbolic objects, SOL (Symbolic Object Language).

Visualization in a Symbolic Table

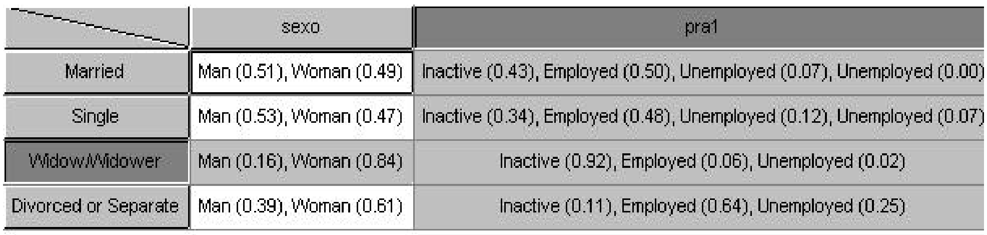
In a symbolic table, rows are symbolic objects and columns are variables. Depending on the type of variables, in each cell will appear distributions, intervals,...

In the following table, we have 4 symbolic objects that describe 4 marital status by means of variables sex, relation to labour activity,...



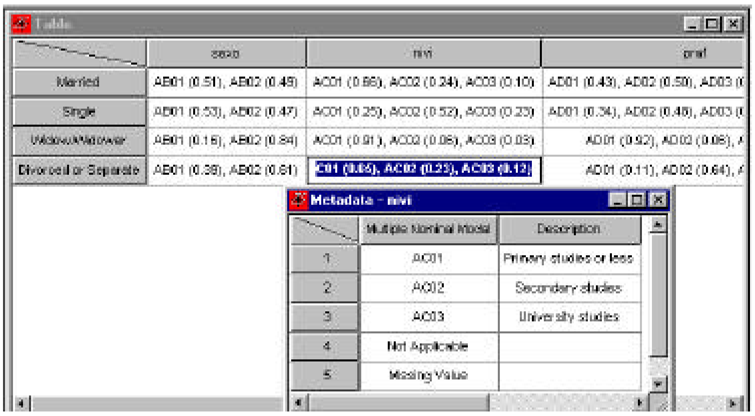
	sexo	pra1
Married	Man (0.51), Woman (0.49)	Inactive (0.43), Employed (0.50), Unemployed (0.07), Unemployed (0.00)
Single	Man (0.53), Woman (0.47)	Inactive (0.34), Employed (0.48), Unemployed (0.12), Unemployed (0.07)
Widow/Widower	Man (0.16), Woman (0.84)	Inactive (0.92), Employed (0.06), Unemployed (0.02)
Divorced or Separate	Man (0.39), Woman (0.61)	Inactive (0.11), Employed (0.64), Unemployed (0.25)

We can select in this table symbolic objects (rows) as well as variables (columns). In this case, we have selected the objects “Widow/Widower” and the variable “Relation to labour activity”.



	sexo	pra1
Married	Man (0.51), Woman (0.49)	Inactive (0.43), Employed (0.50), Unemployed (0.07), Unemployed (0.00)
Single	Man (0.53), Woman (0.47)	Inactive (0.34), Employed (0.48), Unemployed (0.12), Unemployed (0.07)
Widow/Widower	Man (0.16), Woman (0.84)	Inactive (0.92), Employed (0.06), Unemployed (0.02)
Divorced or Separate	Man (0.39), Woman (0.61)	Inactive (0.11), Employed (0.64), Unemployed (0.25)

The category identifiers can be replaced by automatically generated identifiers, and to display a metadata window with the complete labels.



	sexo	nivi	prof
Married	AB01 (0.51), AB02 (0.49)	AC01 (0.66), AC02 (0.24), AC03 (0.10)	AD01 (0.43), AD02 (0.50), AD03 (
Single	AB01 (0.53), AB02 (0.47)	AC01 (0.25), AC02 (0.52), AC03 (0.23)	AD01 (0.34), AD02 (0.48), AD03 (
Widow/Widower	AB01 (0.16), AB02 (0.84)	AC01 (0.91), AC02 (0.06), AC03 (0.03)	AD01 (0.92), AD02 (0.06), A
Divorced or Separate	AB01 (0.39), AB02 (0.61)	C01 (0.65), AC02 (0.23), AC03 (0.12)	AD01 (0.11), AD02 (0.64), A

	Multiple Nominal Modal	Description
1	AC01	Primary studies or less
2	AC02	Secondary studies
3	AC03	University studies
4	Not Applicable	
5	Missing Value	

Zoom Star Visualization

There are two types of zoom star visualization, 2D and 3D, which provide different levels of detail. The 2D representation provides a global impression of the symbolic object, whereas 3D representation provides much more detailed information.

The Zoom Star representation is derived from Kiviat Diagrams where each axis corresponds to a variable. In the same graph we can represent categorical variables, intervals, weighted values, taxonomies,... without overloading the graph.

The following table summarizes the representation of each variable depending on its type.

Variable Type	*Axis Description*
Quantitative	Graduated axis
Categorical	Dots equally distributed on the axis
Categorical not weighted	Axis drawn in black
Categorical weighted	Axis drawn in claret
Not applicable	Axis drawn in grey

The limit for variables to be represented is 24 and for categories is 15.

Selecting an axis with the mouse, we can display the distribution of the chosen variable (histogram). Moreover, we can also display taxonomies and dependencies of a variable by clicking on the icon that appears in the corresponding axis.

Graphics can be moved right, left, up and down for a better visualization.

2D Zoom Star

In the 2D Zoom Star, axes are linked by a line that connects most frequent values of each variable. If there were a tie of the most frequent value in several categories, the line would link all of them.

In the presence of an interval variable, the line is linked to the minimum and maximum limits and the entire area is filled.

For instance, we have defined symbolic objects as groups of population defined by sex, age, marital status, level of education and relation to activity in the P.R.A. survey. We have obtained 314 symbolic objects, which are the combination of the modalities of these variables.

In this case, as we use a survey, the distribution has been calculated taking into account sampling weights.

In the following graph, we can see two mother-daughter variables. Daughter variables that are N.A. appear in the graph as a grey axis. On the right, we can see the distribution of one of the variables.

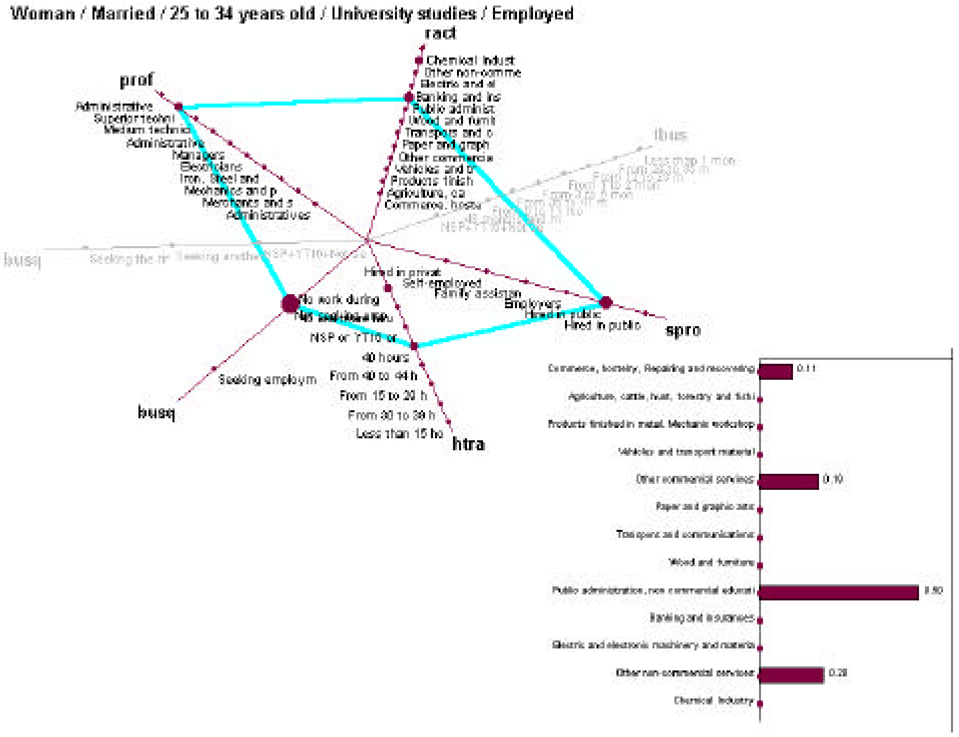


Fig.3: 2D Visualization with Mother-Daughter variables and associated distribution of one of the axis.

3D Zoom Star

In the 3D representation, we can see distributions corresponding to each variable with weighted values. Numerical variables are represented by rectangles from the minimum to the maximum value.

For example, the distribution of the symbolic object "Woman" in the P.R.A survey corresponding to a quarter in Alava is the following,

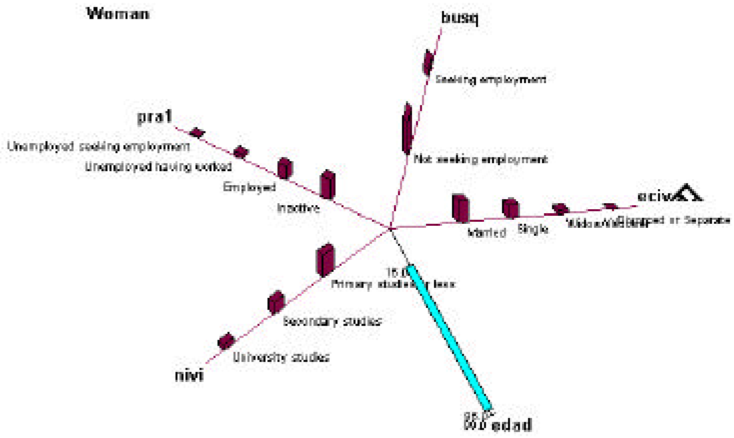


Fig.4: 3D Visualization with interval variable and taxonomy in variable eciv.

SOL Visualization

SOL is a natural language that can be read (and written) easily by users.

Variable Definition

In the assertions file, firstly there is the definition of the variables used to describe the symbolic objects. The information details the type of variable and its categories. Subsequently, rules among variables are included.

```
variable busq1
  nominale {"Seeking employment", "Not seeking employment"}
  multiple,mode=probabilist;

variable busq2
  nominale {"Seeking the first employment", "Seeking other employment (not the
  first one)", "NSP+PY16+ Not seeking employment"}
  multiple,mode=probabilist;

variable tbus1
  real [0:99]
  interval;

busq2 is applicable if busq1 in {"Seeking employment"}
tbus2 is applicable if busq1 in {"Seeking employment"}
```

Symbolic Object Definition

After the definition of variables, all symbolic objects are included in SOL.

Woman / Married / 25 to 34 years old / University studies / Employed =

busq1 = Not seeking employment (1.00)

And busq2 = Not Applicable

And prof2 = Merchants and salespersons (0.11), Administrative managers (0.19), Medium technicians and professionals (0.19), Superior technicians and professionals (0.33), Administrative auxiliaries (0.19)

And ract2 = Commerce, hostelry, Repairing and recovering (0.11), Other commercial services (0.19), Public administration, non commercial education (0.50), Other non-commercial services (0.20)

And tbus2 = Not Applicable

And htra2 = No work during the week (0.28), 40 hours (0.34), From 40 to 44 hours (0.09), From 15 to 29 hours (0.10), From 30 to 39 hours (0.19)

And spro1 = Hired in private sector (0.19), Self-employed (0.11), Hired in public administration (0.62), Hired in public company (0.09)

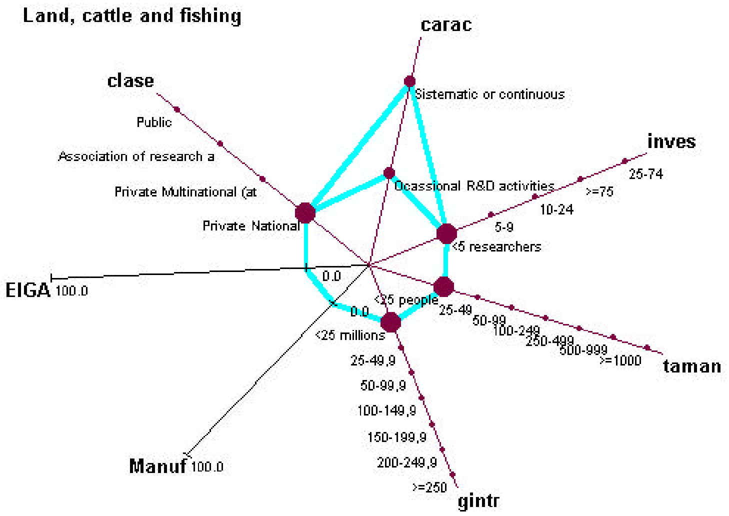
Comparison of several Symbolic Objects

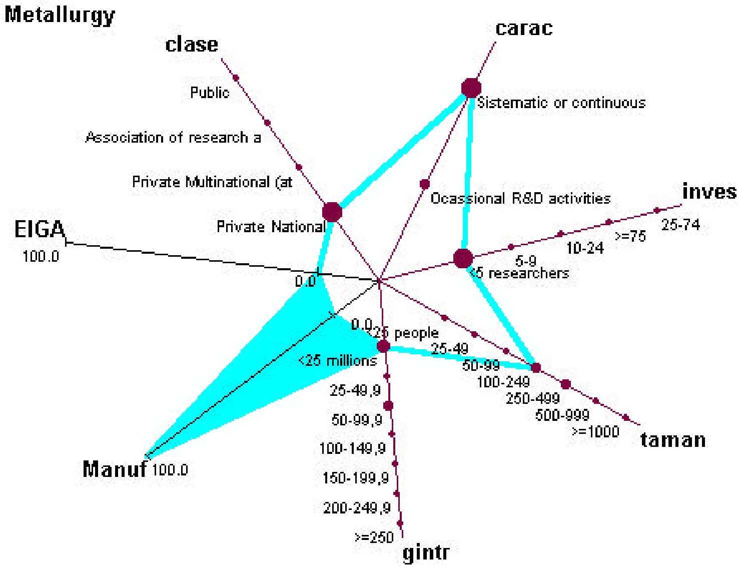
The comparison of several symbolic objects is easier using the 2D representation. We compare if the shapes of the lines that link the axes are similar.

Example

From the survey of Enterprises doing R&D in the Basque Country, we have built some symbolic objects describing branches of economic activity. From the 18 available branches, we have chosen 2 to compare them, “Land, Cattle and Fishing” and “Metallurgy”.

The chosen variables to describe both branches are: Type of Enterprise, Type of R&D activity, number of researchers in the activity, size of the enterprise in staff, intramural expenses, percentage of researching dedicated to manufacturing products and energy.


Land, cattle and fishing

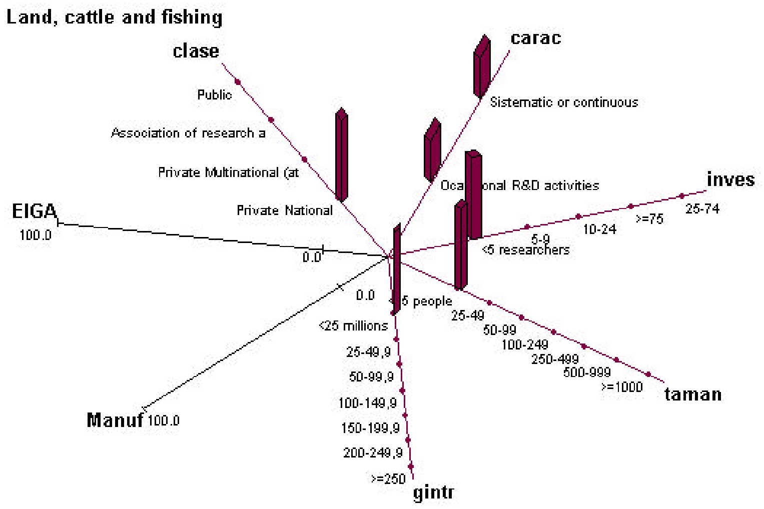

Metallurgy

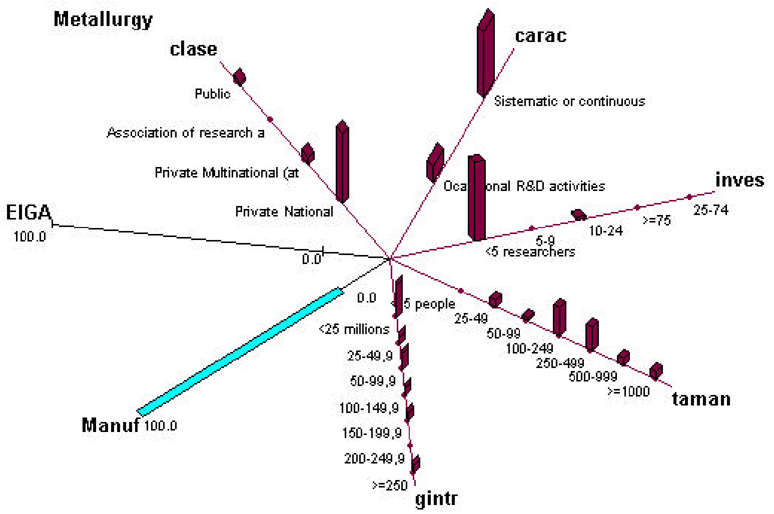
From the graphs we can draw the following conclusions:

The two branches differ in the character of the R&D activities, in “Land, Cattle and Fishing” the activities can be both systematic or occasional, whereas in “Metallurgy” the activities are mostly systematic. Another difference is the size of the enterprises, in the metallurgic industry they are larger than in the “Land, Cattle and Fishing” branch. Moreover, the metallurgic industry uses 100% of intramural expenses for researching manufacturing products.

The comparison with histograms (3D representation) also provides relevant information about the distributions.

We represent the same 3D graphs of branches of economic activity as in the previous example, to obtain more information.





Now, we can observe better the differences between distributions in the two branches. In “Metallurgy”, the distributions of the variables “size of the enterprise” (taman) and “intramural expenses” (gintr) are much more dispersed among all categories, whereas in “Land, Cattle and Fishing” the distributions of these two variables are centred in a unique value.

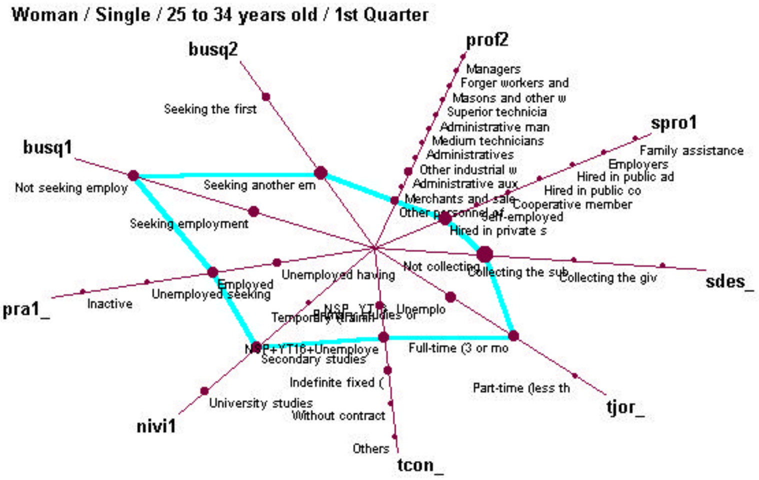
Analysis of Evolution of Symbolic Objects

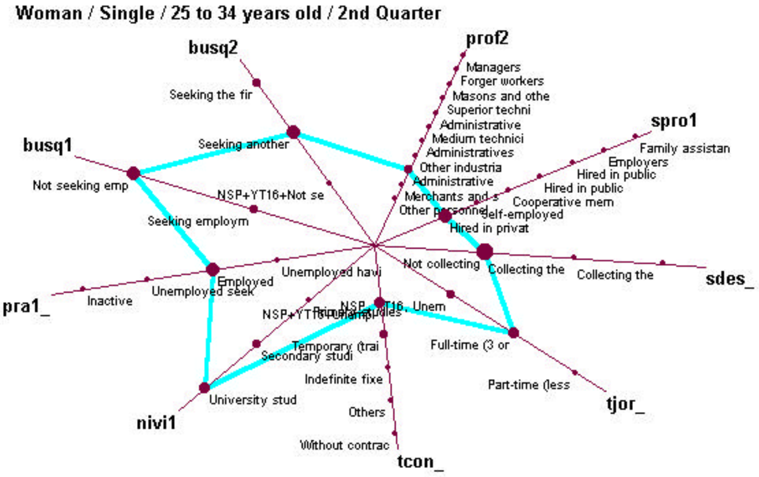
The Zoom Star representation can also be used to analyse the evolution of a symbolic object. The visualization of different versions of the same object would facilitate the identification of stable values or values that show a large variation from one version to another.

Example

Since the P.R.A. survey is a panel survey taken each quarter with a 1/8 rotation, we will study the evolution of the same symbolic object in two consecutive quarters.

We have created symbolic objects crossing the variables sex, marital status and age. The chosen object for the study in the two quarters is “Woman / Single / 25 to 34 years”.





For the shape of the line that links the axes, we can observe that there has been an evolution in the level of education (nivi1) of the group. In the first quarter, most of the women have Secondary studies, whereas in the second period most of them have University studies. Another change occurs in the profession (prof2) where in the first quarter the group belongs mostly to “Other Services personnel” and in the second quarter mostly to “Administrative Auxiliaries”.

On the contrary, there are no significant changes in the variables “Relation to activity”, “Search for employment”, “Professional situation”, “Length of working day” and “Type of contract”.

Advantages of using Symbolic Objects

The Zoom Star visualization of an object allows complex data to be represented with different levels of detail.

This representation allows one object at a time to be visualized, or possibly several objects side by side. This is the main difference with regard to methods which represent points clouds and which seek to interpret interactions between variables.

Chapter
4

Basic Statistics for Symbolic Objects

Basic Statistics of Symbolic Objects consists of a set of graphs and summary measures depending on the type of variable.

If the variables are multinomial, we can draw frequency graphs such as bar graphs and pie charts.

If the variables are interval, we can draw frequency graphs with central tendency and dispersion measures. Moreover, we can represent biplots.

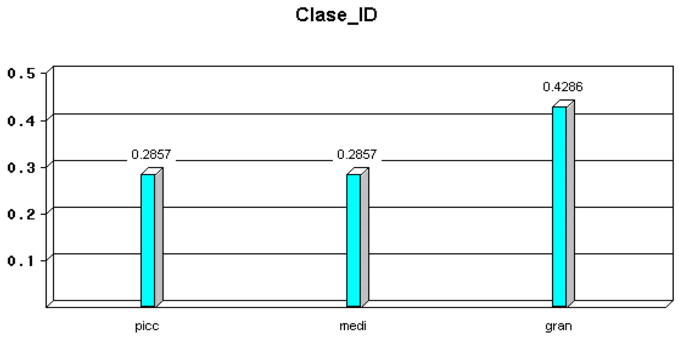
Finally, if the variables are probabilistic, we can draw graphs of capacities.

Frequencies for Multinomial Variables

Bar Graph

These are graphs that represent the distribution of a multinomial variable in a set of symbolic objects.

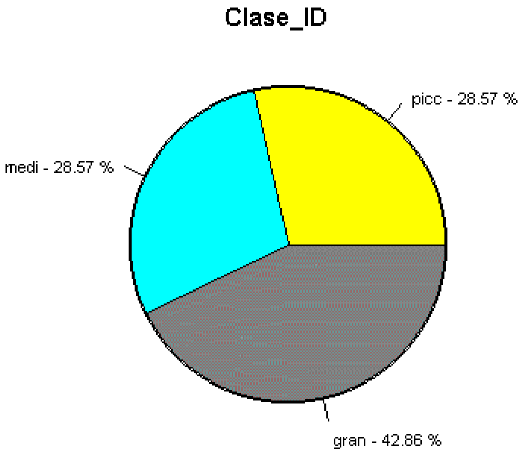
For instance, we have represented the variable Clase_ID with the categories: small (picc), medium (medi) and large (gran). We have measured the frequency of each category in all symbolic objects of a given array. From the graph, we can see that the most frequent category in the array of symbolic objects is “large”.



Relative frequencies (multinominal) - variable 'Clase_ID'

Pie chart

For the same type of variable, we can also draw the distribution in a pie chart.

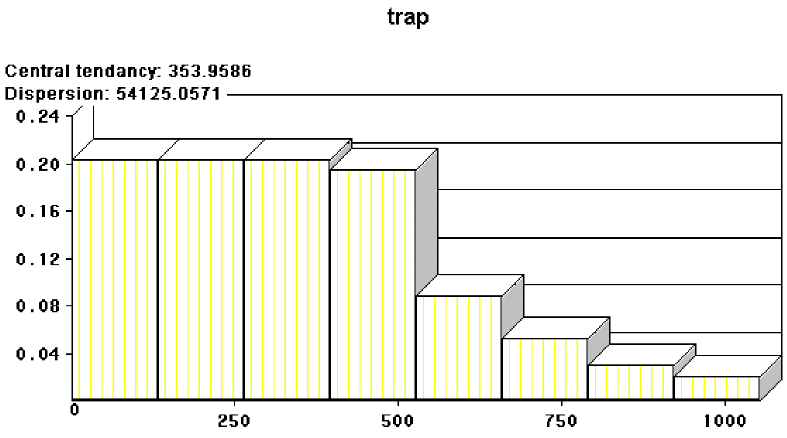


Relative frequencies (multinominal) - variable 'Clase_ID'

Frequencies for Interval Variables

We can study the distribution of a variable building a “symbolic” histogram where the values of the variable are intervals. The number of intervals is chosen by the user and varies from the minimum to the maximum value of the variable chosen.

In the graph, central tendency and dispersion measures are included. In this example the interval variable is "Minutes dedicated to Main Job per day” of the Time Use Survey and 8 classes have been chosen to represent it.

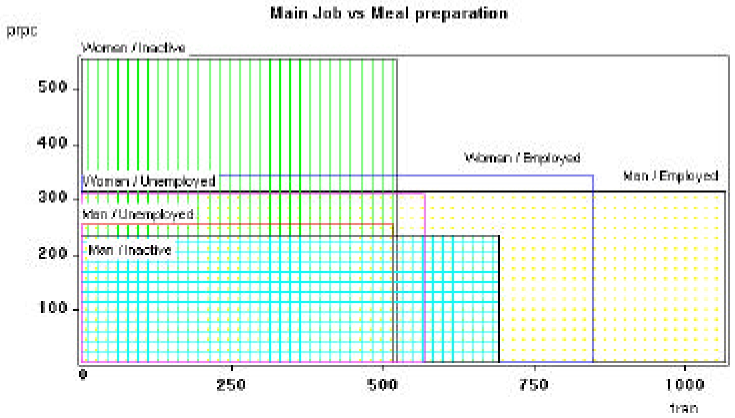


Relative frequencies (interval) - file SEX_PRAT.SDS - variable 'trap' (limits: 0.0 / 1069.0) - 8 classes

Biplot

This is a graph that represents two interval variables. Each object is a rectangle in the plane defined by the two variables.

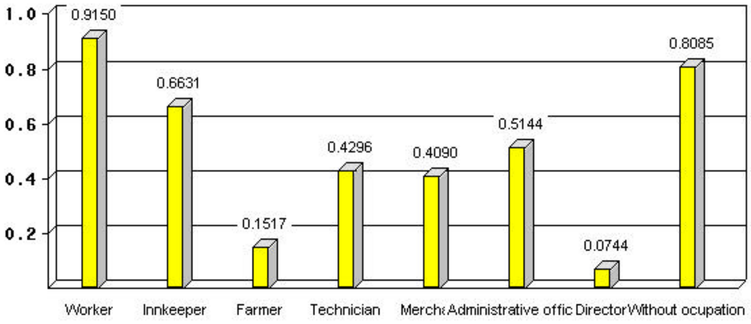
We have drawn two variables from the Use of Time survey: “Time dedicated to Main Job” and “Preparation of meals”, both in minutes. The 6 symbolic objects (rectangles) correspond to groups defined by crossing the variables Sex x Relation to Activity.



From the graph, we can observe that the group that dedicates more time to meal preparation is “Woman Inactive” and less time “Man Inactive”. On the other hand, the group that dedicates more time to Main Job is “Man Employed” and the least is “Man Unemployed”.

Capacities for Probabilistic Variables

This graph represents the capacity of the categories to reach 1 in a set of symbolic objects. In this case, we have represented the variable “Profession” with 8 categories of symbolic objects built from the Use of Time survey.



Capacities - file CAP_ING.SDS - variable 'PROF_ING'

Chapter 5

Conclusions

Traditionally, Official Statistics perform analysis of flat files where the statistical units are individuals. We have proposed new statistical units here called Symbolic Objects, which contain more information than the former.

These new statistical units provide the following advantages in Official Statistics:

a. Statistical Offices from different countries manipulate almost equivalent concepts (such as unemployment data or road accidents), but these concepts are possibly described by some different variables and different official nomenclatures ⇒ Symbolic Objects provide a framework for *describing, unifying and analysing* these heterogeneous concepts.

b. Dissemination of the results of analyses is one of Statistical Offices main tasks ⇒ symbolic objects provide an efficient ergonomic way of *presentation* of data with complex structure.

c. Easiness to transform the data into *array structures* .

d. Symbolic Objects provide a nice way to represent *aggregated and hierarchical data*.

Other uses of Symbolic Objects

In addition to the above mentioned uses, new applications of symbolic objects are under study, such as:

Data Protection

In relation to interval objects, another application will be data protection by means of guarantee intervals. Thus, sensitive data from a relational database will have its privacy assured if instead of giving the punctual datum as a result to a query, the result will be a symbolic object with an interval variable.

Example:

Query: Years in the enterprise and Salary of Mr. Smith?

As the result of the query is confidential, we can return and interval instead of the exact salary.

Mr. Smith = [Years = 3] ∧ Salary = [3.5, 4.2]

Future of the Project

The first stage of the SODAS Project closed with the resulting software SODAS version 1.04. This software has implemented the modules named in the notebook and several others will be explained in successive papers.

A new stage in the project is currently under way, SODAS II, which will improve the results obtained to date and will include new modules to manage and analyse Symbolic Objects.

Chapter 6

References

[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. et al.

Analysis of Symbolic Data. Springer-Verlag (2000).

[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 numerique et symbolique. International Seminar of Statistics in the Basque Country. Volume 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 donnes 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 Nº20821 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.