**ECML / PKDD - 2004 Tutorial Programme**

**Tutorial T2**

# Symbolic Data Analysis

***Pisa, September 20th, 2004***

*Edwin Diday*
Université Paris-IX Dauphine
Paris, France
diday@ceremade.dauphine.fr

*Carlos Marcelo*
Instituto Nacional de Estatística
Lisboa, Portugal
carlos.marcelo@ine.pt

# FROM DATA TO KNOWLEDGE:

# SYMBOLIC DATA ANALYSIS AND THE SODAS SOFTWARE

**E. Diday**

**University Paris 9 Dauphine, Ceremade**
**Place Du Maréchal de Lattre de Tassigny**
**75016 Paris, FRANCE**

**OUTLINE**

# AN INTRODUCTION TO SYMBOLIC DATA ANALYSIS AND THE SODAS SOFTWARE

**E. Diday**

**University Paris 9 Dauphine, Ceremade**
**Place Du Maréchal de Lattre de Tassigny**
**75016 Paris, FRANCE**

***ABSTRACT.*** *In order to take better care from the complex data contained in huge Databases and to extract new knowledge, Symbolic Data Analysis (SDA) get out from standard data tables (imposed by standard statistical or Data Mining software, where n units are described by p categorical or numerical variables). Symbolic Data represented by structured variables, list, intervals, distributions and the like keep better this complexity. In these Databases, first and second order units are distinguished and modeled with standard descriptions(numerical or categorical variables) for the first and with "symbolic description" for the second. Like the units, the theory increases in generality. The objective of SDA is to discover new knowledge (as concepts modeled by "symbolic objects") from Databases and so, to extend Knowledge discovery, Data Mining and Statistics to new kinds of units (like classes, categories or concepts) described by symbolic data. The underlying structure of symbolic objects is defined by Galois Lattices. Finally some practical tools of SDA and the European software SODAS from EUROSTAT are presented.*

*KEYWORDS: Exploratory Data Analysis. Data Mining. Symbolic Data Analysi., Classification, Knowledge discovery. Conceptual lattice. Mixture decomposition.*

## 1) Introduction

A first meaning of a "category" is given in natural language where a "category" is a class of similar units. A second meaning of a "category", is defined in philosophy by the quality which can be attributed to an unit. Hence, Aristotle has defined ten categories: (substance, quantity, relation, quality, place, time, situation, action, behavior, passion) and Kant has defined four classes of categories (quantity, quality, relation, modality).

Databases sometimes huge, are now building and growing up everywhere in order to manage companies or to get decision aid systems. These databases contain sets of units in relation with sets of variables which can be categorical. Categories are considered in this paper from these databases point of view. Hence, categories can be first defined by a value of a categorical variable. It is then easy to associate to each category the set of units of the database which satisfies it. Having this set, it is possible to get a description of this category by using other variables of the database. For instance, if we wish to study the behavior of succer teams and their success (as the number of goals), the units are the teams and not the players. In a "soccer database" , the categorical variable "team" associates to each player its team. Hence, from the category "Italian team" we can obtain the set of players of the Italian

team of the database. We can then model this category by describing this set by other variables as the age, height, salary ... If we consider that each category is a unit of higher level, we obtain a data table where the units are categories described by variables defined on the set of categories instead of the initial variable which was defined on the set of players. The values of these variables have necessarily to take care of the variation of the units inside of each category. For example, by taking care of the variation of the age, height, salary among the set of players of the Italian team we obtain intervals or set of values or distributions. For example, if Tom is an Italian player, we have age (Tom) = 20 for the initial variable "age". But if we consider the category "Italian players" as a unit, we need to use a new kind of variable defined on the categories (here, the teams). If we denote them by capital letters, we obtain AGE (Italian team) = [18, 22] which means that the age varies between 18 and 22. We could choose also as a value of this variable the inter-quartile interval or the distribution of the age of the Italian team's players. These new kinds of variables are called "symbolic variables" and their values are called "symbolic data" as they are not necessarily numbers on which the standard numerical operators can be applied.

The section 2 and 3 are devoted to categories and concepts modelisation. A concept is generally defined by an"intent" and an "extent". The extent of a concept constitutes a category of units which satisfies the properties of its intent. Hence, to any concept it is possible to associate a category. In order to model a concept, we can use symbolic variables able to describe its extent, like for categories, but this is not sufficient as we have also to model its intent. That is why in section 4 we introduce "symbolic objects" which are a way to model concepts including their intent and a way of calculating their extent. In order to build a symbolic object four spaces are needed. From a concept given in a first space called "set of concepts" we obtain its extent among a second set called "set of individuals". These two sets form what we have called the "real world set". The extent of the concept is a category to which we associate a symbolic description in a third set called "the set of description". Finally we build the "symbolic object" associated to the given by adding a matching operator and a membership function to this symbolic description. The set of symbolic objects constitute the fourth set needed to modelise a concept by a symbolic object.

How to extract new knowledge from a set of classe, categories or concepts described by a symbolic description? This is the objective of the so-called "Symbolic Data Analysis" framework. The main principle and the historical origin of SDA are presented in the section 5. In 6.1 we present an historical view on description of classes followed by an overview on the main publications in symbolic data analysis. In 6.3 we present some natural sources of

symbolic data. For example, categories are a natural source of symbolic data, as their description need symbolic data if we wont to keep its internal variation. How to produce categories? Categories can be obtained from the categorical variables of a database or from any Cartesian product of such variable (for example, if the variable "country" contains 5 categories, the Cartesian product country x sex contains 10 categories). Categories can also be obtained from a decision tree in an supervised way or in a supervised way by clustering.

The input and output of a SDA is presented in the section 7. The input are classes, categories, or concepts considered as units modeled by symbolic data sometimes with background knowledge (taxonomic variables and dependencies between variables). The main output are symbolic objects (called SO) extracted for instance, from a clustering or a decision tree extended to symbolic data. From these SO new symbolic descriptions are obtained which can constitute the starting point of a new SDA and so on. In section 8 we present the syntax of a special kind of symbolic objects called "assertions". The way of calculating their extent by using their membership function is presented in section 8.2. In the same section other possible classes of SO are also presented. Some tools concerning symbolic objects (SO) are presented in the section 9, as order and dissimilarities between SO and we give some ideas on how prototypes can be built from a set of concepts or from a set of symbolic descriptions. The underlying structure of SO is studied in section 10 where we show that Galois conceptual lattices generalized to symbolic data constitute an underlying natural structure for SO. This result is illustrated by a simple example. In 10.2 we summarize the main mathematical operators needed in the SDA framework. In 10.3 some statistical aspect for structuring symbolic data are recalled. For instance, a symbolic variable can be considered as a random variable whose value for each second order unit as a concept or a category is a random variable defined on its extent. In the section 10.4, we give some ideas on how to measure the quality, robustness and reliability of SO in the case where we know the extent of a given concept. In section 11 we give the main principle and methods of symbolic data analysis which are or on the way to be included in the SODAS software. Finally, this (free) SODAS software (Bock, Diday (2000)) issued from the work of seventeen European teams and supported by the European Community is summarized in section 12. The input of this software can be a query to a database from which a module called DB2SO creates a symbolic data file on which several symbolic data analysis methods included in SODAS can be applied. In the last issue of SODAS it is also possible to enter directly native symbolic data (i.e. units already described by symbolic data). Finally, in section 13 some advantages in the use of symbolic objects are summarized. Among them, the possibility of propagate a concept from a

database to another database obtained for instance by the same survey in a different time or in a different region.

## 2) Categories

### 2.1 Definition of a category and a categorical set in a database

A categorical variable is a mapping defined on a set Ω whose value for each unit of Ω is a name. A category is an entity defined by the name of its value among the set of possible values taken by a categorical variable in a database. A categorical set C is a set of units having the same value A for a categorical variable. This categorical set C is called the extent of the category A in the database. For example, the category "Senegal team" is a value of the categorical variable "country of the team" defined on a set of international players. The set of Senegal players which are in the database are defined by (country of the team) $^{-1}$("Senegal team") and constitutes a categorical set called the extent of the category "Senegal team".

2.2 Modeling categories by symbolic data: what are "symbolic data"?

In order to understand what is a given category we need to describe it. This description can be obtained from the other variable of the database. As we have seen it in the introduction, standard univalued variables are not sufficient in order to describe categories. Symbolic data can better take care of the internal variation of the values taken by the units which are inside a category. Hence, a category can be modelled by its symbolic description defined by the values taken by several symbolic variables on this category. For example, the category "Senegal team" is a value of the categorical variable "team". Its symbolic description is (AGE (Senegal team), HEIGHT(Senegal team)) if we retain just these two symbolic variable in order to describe this category. It is also possible to add other variables standard or symbolic variables describing the category itself as for example the name of the manager of the team or the correlation between the initial variables (in the example, the correlation between the age and the height in the set of units of the Senegal team).

They are defined in the following way in the SODAS software: "Symbolic data tables" constitute the main input of a Symbolic Data Analysis. They are defined in the following way: columns of the input data table are « symbolic variables » which are used in order to describe a set of units called "individuals". Rows are called « symbolic descriptions » of these individuals because they are not as usual, only vectors of single quantitative or categorical values. Each cell of this « symbolic data table » contains data of different types:

(a) Single quantitative value: for instance, if « height » is a variable and w is an individual : height(w) = 3.5. (b) Single categorical value: for instance, Town(w) = London.

(c) Multivalued: for instance, in the quantitative case, height(w) = {3.5, 2.1, 5} means that the height of w can be either 3.5 or 2.1 or 5. Notice that (a) and (b) are special cases of (c).

(d) Interval: for instance height(w) = [3, 5], which means that the height of w varies in the interval [3, 5].

(e) Multivalued with weights: for instance a histogram or a membership function (notice that (a) and (b) are special cases of (e) when the weights are equal to 1 or 0).

Variables can be: (g) Taxonomic: for instance, « the colour is considered to be "light" if it is "yellow", "white" or "pink" . (h) Hierarchically dependent: for instance, we can describe the kind of computer of a company only if it has a computer, hence the variable “does the company has computers? “ and the variable “ kind of computer” are hierarchically linked.

(i) With logical dependencies, for instance: « if age(w) is less than 2 months then height(w) is less than 10 ».

Many examples of such symbolic data are given in the chapter 3 in Bock, Diday (2000). Figure 1 gives some examples of such data:

| | **WAGES** | **TOWN** | **SOCIO-ECONOMIC GROUP** | |
|---|---|---|---|---|
| **CATEG 1** | **{3.5}** | **{London}** | **{Personal of service}** | |
| **CATEG 2** | **[ 3 , 8 ]** | **{Paris, London}** | | |
| **CATEG 3** | **{3.1 , 4.6, 7.2}** | | **{ 0.1 Manager, 0.6 Manual, ...}** | |
| **CATEG 4** | **[(0.4) [2,3[ , (0.6) [3, 8]]** | | | |

**Figure1. A "symbolic data table": each cell contains an example of "symbolic data".**

### 2.3 Joining categories in databases

For example, the categorical variable "town" (whose values are London, Cambridge, …) defined on a set of inhabitant given by a census, associates its leaving town to each inhabitant. Then, to the categorical value "London", we can associate the categorical set of the inhabitants of London. In another census the variable "town" defined on a set of schools given by another census, associates its leaving town to each school. Then, to the categorical value "London", we can associate the concept: "school of London", whose extent is the set of schools of London. It is then possible to model the category "London schools and people" by joining the symbolic descriptions of the category "people leaving in London" and "school of London" without having to merge both census databases at the level of people or schools. In

section 6 and 7.2 we will see how to induce from the database these symbolic descriptions and symbolic objects by a generalization process.

## 3.Concepts and categorisation

### 3.1 Definition of a concept, its intent and extent

A "concept" is generally defined by a set of properties called "intent" and a set of individuals which satisfy these properties called "extent" or "extension" , this extent can be considered as a "categorical set". By using "universal ideas" for "concepts", Arnault and Nicole (1662) have brilliantly defined these notions in the framework of the "Port-Royal school", in the following way:

*"Now, in these universal ideas there are two things which are important to keep quite distinct: comprehension and extension. I call the comprehension of an idea the attributes which it contains and which cannot be taken away from it without destroying it; thus the comprehension of the idea of a triangle includes, to a superficial*

*extent, figure, three lines, three angles, the equality of these three angles to two right angles, etc. I call the extension of an idea the subjects to which it applies, which are also called the inferiors of a universal term, that being called superior to them. Thus the idea of triangle in general extends to all different kinds of triangles".*

There are two kinds of "concepts":

i) The "concepts of the real world" as a town, a region, a scenario of road accident, a kind of unemployment,... That kind of concept is defined by an "intent" and an "extent" which exists has existed or will exist in the real world.

ii) The "concepts of our mind" (among the so called "mental objects" explained by J.P. Changeux (1983)) which model in our mind concepts of our imagination or concepts of the real world by their properties and a "way of finding their extent" (by using our sensors), and not the extent itself as (for sure!), there is no room, in our mind, for all the possible extents.

A "symbolic object" models a concept in the same way as our mind does it, by using a description "d" (representing its properties) and a mapping "a" able to compute its extent , for instance, the description of what we call a "car" and a way of recognizing that a given entity of the real world is a car. Hence, whereas a concept is defined by an intent and an extent, it is modeled by an intent and a way of finding its extent by "symbolic objects" like those in our mind. Notice that it is quite impossible to obtain all the characteristic properties of a concept and its complete extent. Therefore, a symbolic object is just an approximation of a concept

and the problem of the quality, robustness and reliability of this approximation arise. This important question will be discussed in § 4.

### 3.2 Concepts versus categories and categorization

Hence, a concept is defined by an intent and a way of finding its extent, it is modeled by a symbolic object. A category is defined by the extent of a concept and it is modeled by a symbolic description of this extent.

Categorization is the process of ranging units by categories. Here, the word "category" is taken in its first meaning: a class of similar units. Having a concept, we have the way of finding its extent, so it is easy to obtain the category defined by its extent. Therefore, the question of categorization in practice is transformed in the question of how to get concepts in practice? In other words, how to get a concept from a Database? There are at least two ways. First finding the extent of the concept, from any category of any categorical variable or from any Cartesian product of categorical variables. An other way for obtaining the extent of a new concept is to use a clustering process on a set of units described by a set of variables: each cluster can be considered as an extent of a concept. Having obtained the empirical extent of a concept which can be considered as an "empirical category" the categorization process is complete only if we are able to give a way which allows the automatical calculation of this extent from the intent of the concept. Having its extent, several parameters are needed in order to model a concept First, the choice of the aggregation mapping T which produces a symbolic description $d_C$ of its extent. Second, the choice of the matching mapping R between the description $d_w$ of an individual w and $d_C$. Third, the choice of a membership mapping which allows the allocation of a new individual to a category. These three choices define what we have called a "symbolic object" that we define more precisely below.

## 4. Modeling concepts by "symbolic objects" and some philosophical aspects

### 4.1 Modeling concepts with four spaces: "individuals", "concepts", "description", "symbolic objects"

In figure 2 the "set of individuals" and the "set of concepts" is considered to be in a set called the "real world" , the "modeled world" is the "set of descriptions" which models individuals (or categories of individuals) and the "set of symbolic objects" which models concepts. We start with a "concept" C whose extent denoted Ext(C/Ω') is known in a sample Ω' of individuals. For example, if the concept is "insurance companies", and 30 insurance companies are among a sample Ω' of 1000 companies. Each individual w of the extent of C in

Ω' is described by using the mapping "y" such that y(w) describes the individual w. The concept C is modeled in the set of symbolic objects by the following steps described in figure 2.

i) We generalise the set of descriptions of the individuals of Ext(C/Ω') with the operator T in order to produce the description $d_C$ (which can be a set of Cartesian products of intervals and (or) distributions). In section 7, we give more details on this operator.

ii) The comparison relation R is chosen in relation with the T choice. For instance, if T = ∪ then R = " ⊆ ", it T = ∩ , then R = "⊇ ".

iii) The membership function is then defined by $a_C(w) = [y(w)\ R_C\ d_C]$ which measures the fit or matching between the description y(w) of the unit w and the description $d_C$ of the extent of the concept C in the database. We can then define the symbolic object modeling the concept C by the triple $s = (a_C, R, d_C)$. In section 7.2 we go back to this definition and we illustrate it by several examples.

When we don't have the the extent of the concepts as input, we get them in the following way:

i) A clustering of a sample Ω' by using the description of the individuals produces a set of classes.

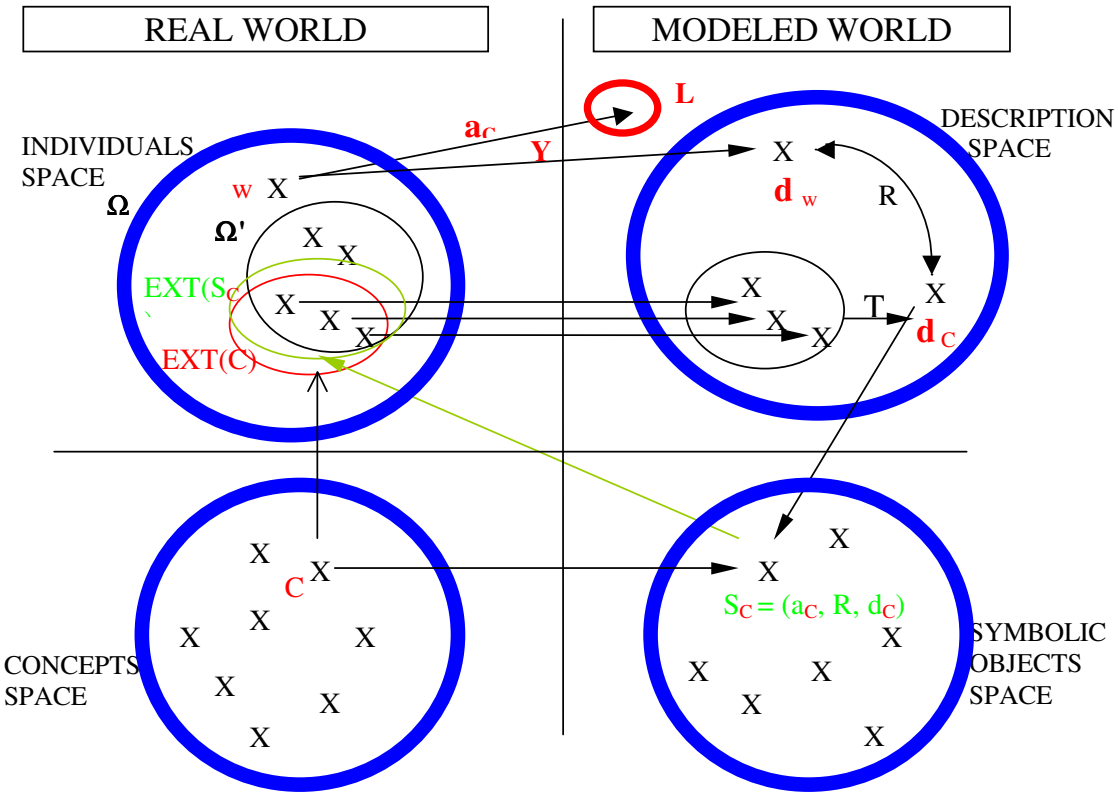



**Figure 2. Modelisation by a symbolic object of a concept known by its extent**

**ii)** To each interesting class denoted A , we associate a concept C and its modelisation by a symbolic object $s_A = (a_A, R_A, d_A)$ with $a_A = [y(w)\ R_A\ d_A]$ where $d_A$ is obtained by using an operator T on the set of the descriptions of the individuals of A, as in the preceding case.

Knowing the extent or empirical category E1 of a concept in Ω' and the extent E2 of a symbolic object which model this concept, the aim of the learning process is to improve the choices of T, R, a, until E1 and E2 be as close as possible. In section 10.4 more details are given on this learning process which yield to the study of the quality, robustness and reliability of a symbolic object.

**4.2 Concepts: four tendencies.**

In the Aristotelian tradition, concepts are characterized by logical conjunction of properties.

In the Adansonian tradition (Adanson (1727-1806) was a pupil of the French naturalist Buffon, very much ahead of his time), a concept is characterized by a set of "similar" individuals. In contrast, with the Aristotelian tradition, where all the members of the extent of a concept are equivalent, a third tendency derived from psychology and cognitive science (see Rosch (1978)), is to consider that concepts must be represented by classes which "tend to become defined in terms of prototyped or prototypical instances that contain the attributes most representative of items inside the class". Wille (1981), following Wagner (1973) says as "in traditional philosophy things for which their intent describes all the properties valid for the individual of their extent are called "concept"".

Symbolic objects combine the advantages of these four tendencies:

- The Aristotelian tradition as they can have the explanatory power of a logical description of the concepts that they represent by "Boolean symbolic objects" (see section 7.2).
- The Adansonian tradition as the members of the extent of a symbolic object are similar in the sense that they must satisfy at best the same properties (not necessarily Boolean, see "modal symbolic objects in section 7.2). In that sense the concepts that they represent are polytheistic.
- The Rosch point of view, as their membership function is able to provide prototypical instances characterised by the most representative attributes. Having a symbolic object $s = (a, R, d_C)$ which model a concept C, prototypes of C are instances w which maximize a(w) in the modal case (see section 7.2). More details on finding prototypes are given in section 9.3.

- The Wagner-Wille property is satisfied by the so-called "complete symbolic objects", which can be proved that they constitute a Galois lattice on symbolic data (see for instance, Diday (1989, 1991, 1998), Brito (1991, 1994), Diday, Emilion (1995, 1997, 2003), Pollaillon, Diday (1997), Pollaillon (1998)).

## 5. Analyzis of the categories by Symbolic Data Analysis

### 5.1 Main principles

In order to analyze categories, their description is needed. As a category is a set of individuals, its description must take care of their internal variation which leads, by using an aggregation mapping, to "symbolic variables" defined on the set of categories. The value of such variable can be a single numerical or categorical value but also, an interval, a set of ordinal or categorical values weighted or not. As they are not purely numerical or categorical they are called "symbolic Data". By using symbolic variables, we can describe a category by a "symbolic description" defined by the values taken by the symbolic variables for this category. Hence, in order to apply the tools of standard Exploratory Data Analysis (clustering, decision trees, factorial analysis, …) to categories we must extent these tools to symbolic data. We can summarize the main principle of a symbolic data analysis by the following steps:

a) A Symbolic Data Analysis needs two levels of units. The first level are individuals, the second level are categories.

b) A category is described empirically by using the description of a class of individuals of its extent.

c) The description of a category must take care on the variation of the individuals of its extent.

d) SDA extends standard data analysis to the case where the units are categories described by symbolic data.

e) The output of some methods of SDA provides new categories (i.e. categories of categories).

f) Next steps: the new categories are considered as new individuals.

### 5.2 Historical and practical origin of the Symbolic Data Analysis field:

The key idea of SDA has been given by Aristotle, four century B. C. . The Aristotle Organon (IV B. C.) clearly distinguishes "first order individuals" (as "a horse" or "a man") considered as a unit associated to an individual of the world, from "second order individuals" (as "the

horse" or "the man") also taken as a unit associated to a class of individuals (the horses, the men).

Our first aim is to extend standard data analysis to second order individuals. For instance, in a census of a country, each household of each region is described by a set of standard numerical or categorical variables (i.e. attributes) given in several relations of a Database. Each household is considered as a "first order individual". In order to study the regions considered as "second order individuals", each variable can be considered as a variable with value a random variable expressing the variations of its values among the households of a region. Hence, we can describe each region in summarizing the values taken by its households, by inter-quartile intervals, or subsets of categorical values, or histograms or probability distributions, etc. depending on the concerned random variable. In such a way, we obtain a "symbolic data table" where each row defines the "description" of a region and each column is associated with a symbolic variable. A simple example of such data is given in section 12.1. An extension of standard Data Analysis to this type of data table is the first aim of what we have called "Symbolic Data Analysis".

Another important aim is to obtain (or "mine") explanatory results (i.e. knowledge) by extracting, the so-called "symbolic objects" which model new "concepts" extracted from the database. A "symbolic object" is defined by its "intent" which contains a way of finding its "extent". For instance, the description of the set of inhabitants of a region and the way of allocating an individual to this region, is called "intent", the set of individuals which satisfy this intent is called "extent". The syntax of symbolic objects must have an explanatory power. For instance, the symbolic object defined by the following expression (see section 7.2, for a formal definition): $a(w) = [age(w) \in [30, 35]\ ] \wedge [\text{Number of children}(w) \leq 2]$, gives the intent of a region of households and at the same time:

i) the description $d = (\ [30, 35]\ , 2)$, where $[30, 35]$ is the inter-quartile interval of the random variable associated to the region for the variable age,

ii) a way of calculating the extent by the mapping "a" defined with the help of the relation $R = (\in, \leq)$.

It means that an individual household "w" satisfies this intent (i.e. belongs to the "extent") of this region, if the age of its members is between 30 and 35 years old and there are less than 2 children.

This very simple kind of symbolic object can be extended at least in the following way: the individuals are of second order (as towns or regions) and represent classes of individuals of first order; therefore the descriptions of the individuals are defined by distributions (the

histogram of the age in a town, for instance). In this case we have to define a different kind: of class description (a "histogram of histograms", for instance), of relation "R" and also, a threshold in order to calculate the extent.

There are several advantages in the use of symbolic objects to model "concepts", one of them, is their ability to be translated into a query of a Database and therefore, to propagate the concepts from a database to another database.

### 5.3 Symbolic Data Analysis is born from which influence?

There was a simultaneous influence of several fields:

- Standard exploratory data analysis (Tuckey (1958), Benzécri (1973), Diday et al (1984) , Saporta (1990), Lebart et al (1998)) where more importance is given to individuals then in standard statistics and where the symbolic approach extend the methods to more complex descriptions of the units and give more explanatory results .
- Artificial Intelligence, Learning Machine (AI) where much efforts has been devoted in finding good languages in order to represent complex knowledge instead of the simple $IR^p$ vectors of the standard statistical units. Notice that the simple language used in order to represent symbolic objects is more inspired from languages based on first order logic ((Michalski (1973), Hayes Roth and McDermot (1977)) then from graph representation (Winston (1979), Sowa (1984)). Notice also, that in symbolic data analysis we are not much interested in the computer language (SQL, C++, JAVA, ...) used in order to represent symbolic objects but much more by their mathematical model, the way of inducing them from the data, their graphical representation, etc.
- Numerical Taxonomy in biology where any species (of insects, mushroom, animals) can be considered as a concept and modeled by a symbolic object.
- Classification in Data Analysis where a class can be modeled by a symbolic object.

## 6. Symbolic Description of classes

### 6.1 Finding a unique description for a class: " T- norm of descriptive generalization"

Let be D the set denoted D of all possible descriptions of the elements of the set of units $\Omega$. Let be $D_C$ the descriptions included in associated to each element of a class of units C (extent of a category or a concept). The T- norm operator is defined on [0, 1]x[0, 1] (Schweizer et Sklar (1983), Bandemer et Nather (1992)). In order to get a symbolic description $d_C$ of C, we use a generalization of the usual T-norm. We call it a " T- norm of descriptive generalization".

This operator is defined on a set of ordered descriptions D where $0_D$ and $1_D$ are respectively the smallest and largest descriptions of D. For example, if D is the set of intervals included in [a, b] and if the order is the inclusion "$\subseteq$" then $1_D$ = [a, b] and $0_D = \varnothing$.

**Definition of a T-norme of descriptive generalization**

T is a mapping $[0_D, 1_D] \times [0, 1] \rightarrow [0_D, 1_D]$ satisfying the following four conditions :

i) $T(u, 1_D) = u$,

ii) monotonicity: $u1 \leq u2 \Rightarrow T(u1,v) \leq T(u2,v)$,

iii) commutativity: $T(u, v) = T(v,u)$,

iv) associativity: $T(u, T(v,w)) = T(T(u, v), w)$.

Due to the commutativity and the associativity of T, we can define a mapping T* from the power set $P(\Omega)$ of the set of units $\Omega$ described in D. If $A = \{a_1,...,a_n\}$ then $T^*(A) = T(a_n, T(a_{n-1}, T(a_{n-2}, T(....., T(a_2, a_1)))$. So we have $T^*(\{a_i, a_j\}) = T(a_i, a_j)$. If we denote $y(A) = \{y(w)/ w \in A\}$ then T*(y(A)) is the description of A for the variable y.

In the same way we can define the T-conorm of descriptive generalization by changing the condition i) $T(u, 1_D) = u$, in i) $T(u, 0_D) = u$.

In Bandemer H., Nather W (1992) many example of t-norm and t-conorm are given which can be generalised to t-norm and t-conorm of descriptive generalization. Among them, it easy to see that the supremum and the infimum are respectively a T-norm and a T-conorm. They are also t-norm and t-conorm of descriptive generalization. It results, that the interval $G(y(C)) = [\inf(D_C), \sup(D_C)]$ constitute a good generalization of $D_C$ as its extent defined by the set of descriptions included in the interval contains C in a good narrow way.

In the case where D is not ordered which happen for instance, when y is a categorical not ordered, we can use the frequency

**Examples of generalisation:**

We give examples with standard quantitative, and of symbolic probabilistic and interval variables as input.

1) y is a standard quantitative variable such that: $y(w_1) = 2.5$, $y(w_2) = 3.6$ , $y(w_3) = 7.1$. $C = \{w_1, w_2, w_3\}$ , D is the set of values included in the interval [1, 100]. Then $G(y(C)) = [2.5, 7.1]$ is the generalisation of $D_C = \{ y(w_1), y(w_2), y(w_3)\} = y(C)$ for the variable y.

2) y is a probabilistic variable of ordered or not ordered categorical variable (where 1(2/3) means that the probability of the category 2 is 2/3), such that: $y(w_1)$ = (2 (1/3), 1(1/3)), $y(w_2)$ = (2(2/3), 1(1/2)), $y(w_3)$ = (2(1/4), 1(1/8)). Then,
G(y(C)) = [[2 (2/3), 1(1/3)], [2 (1/4), 1(1/8)] is the generaliszation of
DC = { y(w1), y(w2), y(w3)} = y(C) for the variable y.
3) y is a variable whose values are intervals such that: $y(w_1)$ = [1.5, 3.2], $y(w_2)$ = [3.6, 4] , $y(w_3)$ = [7.1, 8.4].. C = {$w_1$, $w_2$, $w_3$} , D is the set of intervals included in the interval [1, 100], Then T*(y(C)) = [1.5, 8.4] is the generalisation of $D_C$ = { $y(w_1)$, $y(w_2)$, $y(w_3)$} = y(C) for the variable y.

Instead of describing a class by its t-norm and conorm, many alternatives are possible by taking care, for instance, of the existence of outlier. A good strategy consists of reducing the size of the boundaries in order to reduce the number of outliers. This is done in DB2SO inside the SODAS software (see the appendix). Another choice of DB2SO is to use the frequencies in the case of a not ordered categorical variable.

**Example:**

y is a standard not ordered categorical variable such that: $y(w_1)$ = 2, $y(w_2)$ =2 , $y(w_3)$ =1, C = {$w_1$, $w_2$, $w_3$} , D is the set of probabilities on the values 1,2. Then, G'(y(C)) = [1(1/3), 2(2/3)] is the generalisation of $D_C$ = { $y(w_1)$, $y(w_2)$, $y(w_3)$} = y(C) for the variable y.

We can notice that we obtain an equivalent result by using G(y(C)) = [inf($D_C$ ), sup($D_C$)] if each unit is considered to be an element of the power set P(Ω) of Ω described by the frequencies of the categories. As Ω and the empty set belongs in P(Ω) we get G(y(C)) = [[1(0), 2(0)], [1(1/3), 2(2/3)]].

### 6.2 Finding a disjunction of descriptions for a class

Instead of representing a class C of units by a unique description, it can be interesting to decompose this class in subsets which description is obtained by generalization. The description of C is then the disjunction of these descriptions. This kind of description avoid over generalization.

**Historically: three tendencies** The first proposed by A. de Jussieu (1748) is in the Aristotelian tradition and consists in defining top down the classes by a good choice of the properties which characterize them from the most general to the most specific. In that way we obtain a decision tree where each node is characterized by a conjunction of properties. Many

others have continued this tendency. In the supervised case, by starting from individuals of first order: Belson (1959), Morgan and Sonquist (A.I.D. program (1963)), Lance and Williams (1967), Breiman and al. (1984), Quinlan (1986). By starting from individuals of second order: Pankurst (1978), Payne (1975), Gower (1975), J. Lebbe, R. Vigne (1991), H. Ralambondrainy (1991), Ganascia (1991), Conruyt (1993) in the case of structured data, A. Ciampi, E. Diday, J. Lebbe, E. Périnel, R. Vigne (1995, 2000), Périnel (2000), Aboa (2001), Diday, Bock (2000). In the unsupervised case, Chavent (1998), Diday, Bock (2000). Simultaneously an unsupervised and a supervised approach (see Vrac, Diday (2000), Limam, Diday, Winsberg (2004)).

The second tendency concerns mainly the unsupervised case, put forward by Adanson (1757) who gave the first "Sequential Agglomerative Hierarchical Clustering" (SAHC) algorithm. This well-known "bottom up" algorithm, starting by classes reduced to individuals, merges at each stage the most "similar" classes. This tendency is well represented by Ward (1963), Lerman (1970), Jardine and Sibson (1971), Sneath and Sokhal (1973), Jambu (1978), Roux (1985), Bock (1974), Celeux, Diday, Govaert, Lechevallier and Ralambondrainy (1989), etc. The classes obtained in this way contain similar objects. It is then possible to generalize them in terms of disjunction of conjunction of properties. Whereas, the first tendency yields monotheistic classes by a top-down process, the second produces polytheistic classes by a "bottom up" process. In this framework, a family of methods called "Conceptual Clustering" has been developed in the eighties such as Langley and Sages (1984), Lebowitz (1983), Fisher D.H. (1987), Fisher and Langley (1986) for a review. Instead of producing trees, in Diday (1984), Bertrand (1986) for instance, an ascending process building a pyramid (a generalization of hierarchical trees, allowing overlapping clusters) of polytheistic classes is described. In Brito and Diday (1991), Brito (1994), E. Diday, Rodriguez, Brito (2000), Rodriguez (2000) an ascending pyramid produces monotheistic classes.

The third tendency consists in looking directly for classes and their protype representation. For instance, the "Dynamic Clustering Method" (Diday (1971), Diday and al (1979)), Diday and Simon (1976)), defines a general framework and algorithms which aim to discover simultaneously classes and their protype representation in such a way that they "fit" together as well as possible. In the case of standard data, when the prototypes are means we obtain the standard K-means algorithm. The Dynamic Clustering Method (which can be called also a K-prototype method) has been used with several kinds of inter-class structure (partitions, hierarchies, ...) and prototype representation modes for each class (seeds, probability laws, factorial axis, regressions,...). In Diday (1976), a logical representation of clusters is proposed.

With regards to the "Conceptual Clustering" algorithm based on the Dynamic Clustering Method or inspired by it, mention should be made of Diday, Govaert, Lechevallier, Sidi (1980), Michalski, Diday, Step (1982), Michalski, Stepp (1983) among other pioneers papers in "Conceptual Clustering".

**6.3 The Symbolic Data Analysis field:**

Since the first papers announcing the main principles of Symbolic Data Analysis ((Diday (1987) a, (1987) b, (1989)) much work has been done up to the most recent book published by Bock, Diday (2000) and the proceedings of IFCS'2000, published by Kiers and al (2000) which contains a large chapter devoted to this field. In factorial analysis, P. Cazes, A. Chouakria, E. Diday, Y. Schechtman (1997)) have defined a principal component analysis of individuals described by a vector of numerical intervals and in the same direction R. Verde, F.A.T. De Carvalho (1998) by taking care on given dependence rules, see also Lauro, Palumbo (1998). In the case where the individuals are described by symbolic data, Conruyt (1993) in the case of structured data, A. Ciampi, E. Diday, J. Lebbe, E. Périnel, R. Vigne (1995, 2000), Périnel (2000), Aboa (2001), have developed an extension of standard decision trees. In the same direction E. Perinel and Y. Lechevallier on "symbolic discrimination rules" , M.C. Bravo, J.M. Garcia-Santesmases (1998) on "segmentation trees for stratified data" and J.P. Rasson and S. Lissoir(1998) starting from a dissimilarity between symbolic descriptions have a chapter in Bock , Diday (2000). See also E. Auriol (1995) for a link with the domain of "Case Based Reasoning". In order to select the symbolic variables which distinguish at the best the individuals or classes of individuals, several works have been done such as R. Vignes (1991) and more recently Ziani (1996). It is often useful to calculate dissimilarities between symbolic objects; in that direction mention should be made of C. Gowda and E. Diday (1992), De Carvalho (1994, 1998 a). If each cell of the data table is a random variable represented by a histogram (for instance, the histogram of the inhabitant age of a town), a histogram of histogram can be calculated for instance, by taking care of rules between the variables values in De Carvalho (1998) b, or by using the capacity theory in Diday, Emilion ((1995, 1997), Diday, Emilion, Hillali (1996). Noirhomme and Rouard (1998) give a way of representing multidimensional symbolic data (see chapter 7 in Bock, Diday (2000)), see also E. Gigout (1998) .

Starting from standard data, Gettler-Summa (1992), Smadhi (1995) have proposed a way for extracting symbolic objects from a factorial analysis; in order to extract symbolic objects from a partition, see Stephan, Hébrail, Lechevallier (see chapter 5 in Bock, Diday (2000)) Gettler-Summa M. (1997) and M. Vrac, E. Diday, S. Winsberg, M.M. Limam (2002). Starting

from time-series, Ferraris, Gettler-Summa, C. Pardoux, H. Tong (1995), have defined a way for providing symbolic objects (see chapter 12 in Bock , Diday (2000)) .

More recently, several dissertations have been presented in the Paris 9 - Dauphine University. Mfoumoune (1998), Oldemar (2001) for the sequential building of a pyramid where each node is associated to a symbolic object. Chavent (1998), in order to build a partition of a set of symbolic objects by a top-down algorithm which also provides a symbolic object associated to each obtained class (see chapter 11 in Bock, Diday (2000)), Stéphan (1998) for extracting symbolic objects from a database (see chapter 5 in Bock, Diday (2000)), Hillali (1998) for describing classes of individuals described by a vector of probability distributions, Pollaillon (1998), for extending Galois lattices and extracted pyramid to symbolic data at input and "complete" symbolic objects at output (see section 11.4 in Bock , Diday (2000)) and chapter 1 in E. Diday, Y. Kodratoff, P. Brito, M. Moulet (2000), Tang (1998) for extending Factorial Correspondence Analysis and Rodriguez (2000) for extending regression and Multidimensional Scaling to interval data. De Régnies (2002) for extending Dynamic Clustering to symbolic data. Diday (2002), Vrac (2002) to extend Mixture decomposition to distributions. Synthesis on Symbolic data Analysis can be found in Diday (2000) and Billard, Diday (2003).

### 6.4 Sources of Symbolic Data

Many examples of symbolic data can be found on the same web site.

Symbolic data are generated from any database which contains descriptive categorical variables. For instance in a survey on households, these categories can be simply the regions or in a more complex way, the socio-professional categories (SPC) crossed with categories of age (A) and regions (R). Hence, in this last case, we obtain a new categorical variable of cardinality $|SPC| \times |A| \times |R|$, where $|X|$ is the cardinality of X. The descriptive variables of the households can then be used in order to describe these categories by symbolic data.

Symbolic Data can also appear after a clustering in order to describe in an explanatory way (by using the initial variables) the obtained clusters.

Symbolic data may also be "native" in the sense that they result from expert knowledge (scenario of traffic accidents, type of emigration, species of insects, ...), from the probability distribution , the percentiles or the range of any random variable associated to each cell of a stochastic data table, from time series (in representing each time serie by the histogram of its values or in describing intervals of time), from confidential data (in order to hide the initial data by less accuracy), etc. They result also, from Relational Databases, in order to study a set

of units whose description needs the merging of several relations as is shown in the following example.

**Example:** We have two relations of a Relational Database defined as follows. The first one called "delivery" is given in table 1. It describes five types of deliveries characterised by the name of the supplier, its company and the town from where the supplying is coming.

| Delivery | Supplier | Company | Town |
|---|---|---|---|
| Liv1 | F1 | CNET | Paris |
| Liv2 | F2 | MATRA | Toulouse |
| Liv3 | F3 | EDF | Clamart |
| Liv4 | F1 | CNET | Lannion |
| Liv5 | F3 | EDF | Clamart |

Table 1 Relation "Delivery"

The supplying are described by the relation "Supplying" defined in the following table 2.

| Supplying | Supplier | Town |
|---|---|---|
| FT1 | F1 | Paris |
| FT2 | F2 | Toulouse |
| FT3 | F1 | Lannion |
| FT4 | F3 | Clamart |
| FT5 | F3 | Clamart |

Table 2: Relation "Supplying"

From these two relations we can deduce the following data table 3, which describes each supplier by his company, his supplying and their providing towns:

| Supplier | Company | Supplying | Town |
|---|---|---|---|
| F1 | CNET | FT1, FT3 | ½ Paris, ½ Lannion |
| F2 | MATRA | FT2 | Toulouse |
| F3 | EDF | FT4, FT5 | Clamart |

Table 3: Relation "Supplier" obtained by merging the relations "Delivery" and "Supplying".

Hence, we can see that in order to study a set of suppliers described by the variables associated with the two first relations we are naturally required to take in account the four following conditions which characterise symbolic data:

i) Multivalued: this happen when the variables "Supplying" and "Town" have several values as shown in the table 3.

ii) Multivalued with weights: this is the case for the towns of the supplier F1. The weights ½ means that the town of the supplier F1 is Paris or Lannion with a frequency equal to ½.

iii) Rules: some rules have to be given as input in addition to the data table 3. For instance, "if the town is Paris and the supplier is CNET, then the supplying is FT1.

iv) Taxonomy: by using regions we can replace for instance {Paris, Clamart} by " Parisian Region ".

## 7. Main input and output of Symbolic Data Analysis algorithms

### 7.1 Main input

The input of a Symbolic Data Analysis algorithm is a Symbolic data table (native or obtained by DB2SO in SODAS) from a query to the Database or with some background knowledge as rules or taxonomies as explained in the section 2.2.

### 7.2 Main output: Symbolic objects

Most of the symbolic data analysis algorithms give in their output the symbolic description "d" of a class of individuals by using a "generalisation" process. By starting with this description, symbolic objects give a way, to find at least, the individuals of this class.

**Example:**

The age of two individuals w1, w2 are age(w1) = 30, age(w2 ) = 35, the description of the class C = {w1, w2} obtained by a generalisation process can be [30, 35]. The extent of this description contains at least w1 and w2 but may contain other individuals. In this simple case the symbolic object "s" is defined by a triple: s = (a, R, d) where d = [30, 35], R = " $\in$ " and "a" is the mapping: $\Omega \rightarrow$ {true, false} such that a(w) = "the true value of "age(w) R d" " denoted [age(w) R d]. An individual w is in the extent of s iff a(w) = true.

More formally (see figure 2), let $\Omega$ be a set of individuals, D a set containing descriptions of individuals or of a class of individuals, « y » a mapping defined from $\Omega$ into D which associates to each w $\in$ $\Omega$ a description d $\in$ D from a given symbolic data table. We denote by R, a relation

defined on D. It is defined by a subset E of D×D. If (x, y)∈E we say that x and y are connected by R and this is denoted by x R y . More generally we say that x R y take its value in a set L. We can have L ={true, false}, in this case [d' R d] = true means that there is a connection between d and d'. We can also have L= [0, 1] if d is more or less connected to d'. In this case, [d' R d] can be interpreted as the "true value" of x R y or " the degree to which d' is in relation R with d" (see in Bandemer and Nather (1992), the section 5.2 on fuzzy relations).

For instance, R ∈ {=, ≡, ≤, ⊆ } or is an implication, a kind of matching taking care of missing values, etc. R can also use a logical combination of such operators.

A « symbolic object » is defined by a description "d" , a relation "R" for comparing d to the description of an individual and a mapping "a" called "membership function". More formally:

**Definition of a symbolic object**

*A symbolic object is a triple s = (a, R, d) where R is a relation between descriptions, d is a description and "a" is a mapping defined from Ω in L depending on R and d.*

Symbolic Data Analysis concerns usually classes of symbolic objects where R is fixed, "d" varies among a finite set of coherent descriptions and "a" is such that: a(w) = [y(w) R d] which is by definition the result of the comparison of the description of the individual w to d. More generally, many other cases can be considered. If for instance, the mapping "a" is of the following kind: a(w) = [ $h_e$ (y(w)) $h_j$ (R) $h_i$ (d)] where the mappings $h_e$ , $h_j$ and $h_i$ are "filters" which will be discussed hereunder. There are two kinds of symbolic objects:

- « Boolean symbolic objects » if [y(w) R d] ∈ L = {true, false}. In this case, if y(w) = ($y_1$,...,$y_p$), the $y_i$ are of type (a) to (d), defined in section 2.2.

**Example:**

Let be a(w) = [y(w) R d] with R: [ d' R d ] = $\vee_{i=1,2}$ [ $d'_i$ $R_i$ $d_i$ ] where ∨ has the standard logical meaning and $R_i$ = ⊆ . If y(w) = (colour(w), height(w)), d = ({red, blue, yellow}, [10,15] ) = ($d_1$, $d_2$), colour(u) = {red, yellow}, height(u) = {21}, then

a(u) = [colour(u) ⊆ {red, blue, yellow}]∨ [height(u) ⊆ [10,15]]= true ∨ false = true.

- « Modal symbolic objects » if [y(w) R d] ∈ L = [0,1].

**Example:**

Let be a(u) = [y(u) R d] where for instance R: [ d' R d ] = Max $_{i=1,2}$ [ d'$_i$ R$_i$ d$_i$ ]. The choice of the Max is among many other possible choices related to copulas theory (see Diday (2000)). The "matching" of two probability distributions is defined for two discrete probability distributions d'$_i$ = r and d$_i$ = q of k values by: r R$_i$ q = $\sum_{j=1,k}$ r$_j$ q$_j$ e $^{(r_j - \min(r_j, q_j))}$. By analogy with the Boolean case we denote [ d' R d ]= $\vee^*_{i=1,2}$ [ d'$_i$ R$_i$ d$_i$ ] where $\vee^*$ = Max . With these definitions it is possible to calculate the mapping "a" of a symbolic object s = (a, R, d) where SPC means « socio-professional-category » and d = ({(0.2)12, (0.8)[20 ,28]}, {(0.4)employee, (0.6)worker}) by:

a(u) = [age(u) R$_1${(0.2)12, (0.8) [20 ,28]}] $\vee^*$ [SPC(u) R$_2${(0.4)employee, (0.6)worker}]

Notice that in this example the weights (0.2), (0.8), (0.4), (0.6) represent frequencies but more generally other kinds of weights may be used as "possibilities", "necessities", "capacities", etc. Notice that the Ri depends on this choice, (see Diday (1995), for instance).

## 8. Syntax and properties of symbolic objects

### 8.1 Syntax of symbolic objects in the case of "assertions":

If the initial data table contains p variables we denote y(w) = (y$_1$(w),..., y$_p$ (w)), D = (D$_1$,...,D$_p$), d ∈ D: d = (d$_1$,..., d$_p$) and R' = (R$_1$,...,R$_p$) where R$_i$ is a relation defined on D$_i$. We call « assertion » a special case of a symbolic object defined by s = (a, R, d) where R is defined by

[ d' R d ] = $\wedge_{i=1,p}$ [ d'$_i$ R$_i$ d$_i$ ] where "$\wedge$" has the standard logical meaning and "a" is defined by: a(w) = [ y(w) R d] in the Boolean case. Notice that considering the expression a(w) = $\wedge_{i=1,p}$ [ y$_i$ (w) R$_i$ d$_i$ ] we are able to define the symbolic object s = (a, R, d). Hence, we can say that this explanatory expression defines a symbolic object called "assertion".

For example, a Boolean assertion is:

a(w) = [age(w) $\subseteq$ {12, 20 ,28}] $\wedge$ [SPC(w) $\subseteq$ {employee, worker}]. If the individual u is described in the original symbolic data table by age(u)={12, 20} and SPC (u) = {employee } then: a(u) = [{12, 20 }$\subseteq$ {12, 20 ,28}] $\wedge$ [{employee}$\subseteq${employee, worker}]= true.

In the modal case, the variables are multivalued and weighted, an example is given by

a(u) = [y(u) R d ] with [d' R d ] = f({[y$_i$(w) R$_i$ d$_i$]}$_{i=1,p}$) where for instance,

f({[y$_i$(w) R$_i$ d$_i$]}$_{i=1,p}$) = $\Pi_{i=1,2}$ [ d'$_i$ R$_i$ d$_i$ ] where in case of probability distributions, the "matching" is defined for two discrete density distributions d'$_i$ = r = (r$_1$, ...,r$_k$) and

$d_i = q = (q_1, \ldots, q_k)$ of k values by: $r\ R_i\ q = \sum_{j=1,k} r_j\ q_j\ e^{(r_j - \min(r_j, q_j))}$.

By analogy with the Boolean case we denote $[\ d'\ R\ d\ ] = \wedge^*_{i=1,2}\ p_i\ [\ d'_i\ R_i\ d_i\ ]$ where the meaning of "$\wedge^*$" is given by the definition of the mapping "f". For instance, with these choices, a modal assertion s = (a, R, d) is completely defined by the equality:

$a(w) = [age(w)\ R_1\ \{(0.2)12, (0.8)\ [20, 28]\}]\ \wedge^*\ [SPC(w)\ R_2\ \{(0.4)employee, (0.6)worker\}]$

### 8.2 Extent of a symbolic object s and other possible classes of symbolic objects

In the Boolean case, the extent of a symbolic object is denoted Ext(s) and defined by the extent of a, which is: $Extent(a) = \{w \in \Omega / a(w) = true\}$. In the modal case, given a threshold $\alpha$, it is defined by $Ext_\alpha(s) = Extent_\alpha(a) = \{w \in \Omega / a(w) \geq \alpha\}$.

If for instance the mapping "a" is of the following kind: $a(w) = [h_e(y(w))\ h_j(R)\ h_i(d)]$, different classes of symbolic objects may be defined depending on the choice of $h_e$, $h_e$ and $h_i$. In practice, these mappings may be used for instance, in the following way: $h_e$ is a filter of the extension of the symbolic object; $h_j$ is a filter of the descriptive variables and $h_i$ is a filter on the descriptions. More details may be found in Diday (1998) and in chapter 3 of Bock, Diday (2000). The following example illustrates a kind of filter.

**Example of filter on the extension:**

We associate to each town a symbolic object defined by $a(w) = [\ h_e\ (y(w))\ R\ d]$ where "d" is the description of its inhabitant by using for instance, the histogram associated to each variable (as the histogram of the age). In order that the extension of such symbolic object contains only members of its associated town, the mapping $h_e$ is defined in the following way: $h_e\ (y(w)) = y(w)$ if w is member of the town and if not $h_e\ (y(w)) = HS$ where HS is a dummy value such that $[\ h_e\ (y(w))\ R\ d] = 0$ for any description d.

## 9 Tools for symbolic objects

### 9.1 Order between symbolic objects

If r is a given order on D, then the induced order on the set of symbolic objects denoted by $r_s$ is defined by : $s_1\ r_s\ s_2$ iff $d_1\ r\ d_2$.

If R is such that $[d\ R\ d'] =$ true implies $d\ r\ d'$, then $Ext(s1) \subseteq Ext(s2)$ if $s1\ r_s\ s2$ . If R is such that $[d\ R\ d'] =$ true implies $d'\ r\ d$ then $Ext(s2) \subseteq Ext(s1)$ if $s1\ r_s\ s2$.

**9.2 Dissimilarities and Symbolic description of a class**

Tools between symbolic objects (Diday (1995)) are needed such as dissimilarities (F. de Carvalho (1998), Esposito et al (1998)), matching, merging by generalisation where a t-norm or a t-conorm (Schweizer, Sklar (1983) and Diday, Emilion (1995), (1997)) denoted T can be used, splitting by specialisation (Ciampi et al. (1995)). A recent synthesis and other tools can be found in Bock, Diday (2000). Hausdor-type dissimilarity measures are good example of dissimilarity between symbolic data. Let A and B be two subsets of a set E for which we know the dissimilarity between the units. Then a Haussdorf- type dissimilarity measure can be defined in the follow way $d(A, B) = \text{Max} \{\text{Max}_{a \in A} \text{Min}_{b \in B} d(a, b), \text{Max}_{b \in B} \text{Min}_{a \in A} d(a, b))$, In the case of several variables the mean of these dissimilarities for each variable can be calculated.

It is possible to build an operator T which produces a set of symbolic objects whose extent covers a given class A , the extent of each one covering partially or completly the class A. In that way, see for instance M. Chavent (1997) by a top-down clustering tree or Brito and Diday (1991) in a bottom-up clustering pyramid in the unsupervised case and E. Perinel (1996)) or Gettler-Summa (1995) in the supervised case. A synthesis is given in Bock, Diday (2000). It is also interesting to describe A by a set of symbolic objects which satisfies simultaneously an unsupervised and a supervised criteria (see Vrac, Diday (2000), Limam, Diday, Winsberg (2004)). For instance, in a top-down clustering tree where at each step a spliting variable is choosen which cuts A in two subclasses and optimises a given criterion. This criterion can express (in its unsupervised part), the sum of the two by two dissimilarities of the individuals of each subclass and simultaneously (for its supervised part), the Gini impurity criterion of this class. For instance, being 50 years old can be homogeneous for a class but not discriminant whereas being less than 50 years old can be discriminant of a class but not homogeneous. The process stops when it doesn't improve the criterion and in the final tree, we associate easily to each terminal subclass a symbolic object by the conjunction of the values of the spliting variables used in the branches of the path which defines this subclass. The extent of each obtained symbolic object covers partially A and together they cover A. This method can be applied iteratively to each classe of a partition or of a covering of the intial set of individuals in order to summerize it by symbolic objects. For the supervised part of the criterion, the variable to discriminate can be defined by two categories: the given class and its complementary.

### 9.3 Finding prototypes from a class

If we wish to obtain prototypes of a given class which is member of the set of units, we can use a given dissimilarity (of Hausdorf-type for instance) between any unit and each member of the class. The prototypes are then the units which minimise the sum of these dissimilarities. In the case of a class of concepts the best prototypes are the one which maximise the membership functions of the symbolic objects associated to each concept of the class.

If we consider that a prototype is not necessary a member of the set of units, then the prototypes can be calculated by various ways depending on the variable type. For example, if a symbolique variable Y is of interval type (i.e. Y(w) is an interval), a prototype can be defined in the following way: let M be the mean of the means of the intervals associated to the members of the class and $\Delta$ be the mean of the range of each interval of the class, then the prototype interval can be: $[M- \Delta/2, \quad M+\Delta/2]$. Many other possibilities exists, for instance by using the mean of the min and the mean of the max of each interval of the class. In the case of a variable whose values are subsets of a given set, the probability or the conditional probability (or their product) for each value which appear in the class can be used in order to define a prototype. In the case of a variable whose values are distributions their mean can be used or their distribution (see Diday (2002)).

## 10.Underlying structures of symbolic objects

### 10.1 A generalized conceptual lattice

Under some assumptions on the choice of R and T (for instance $T \equiv Max$ if $R \equiv \leq$ and $T \equiv Min$ if $R \equiv \geq$) it can be shown that the underlying structure of a set of symbolic objects is a Galois lattice (Diday (1989, 1991), Brito (1991, 1994), Diday, Emilion (1995), (1997), Polaillon, Diday (1997), Polaillon (1998), Bock, Diday (2000)), where the vertices are closed sets defined below by « complete symbolic objects ». More precisely, the associated Galois correspondence is defined by two mappings F and G:

- F: from $P(\Omega)$ (the power set of $\Omega$) into S (the set of symbolic objects) such that $F(C) = s$ where $s = (a, R, d)$ is defined by $d = T_{c\in C}\, y(c)$ and so $a(w) = [y(w)\ R\ T_{c\in C}\, y(c)]$, for a given R. For example, if $T_{c\in C}\, y(c) = \cup_{c\in C}\, y(c)$ , $R \equiv « \subseteq »$, $y(u) = \{pink, blue\}$, $C = \{c, c'\}$, $y(c) = \{pink, red\}$, $y(c') = \{blue, red\}$, then $a(u) = [y(w)\ R\ T_{c\in C}\, y(c)] = [\{pink, blue\} \subseteq \{pink, red\} \cup \{blue, red\}) = \{pink, red, blue\}] = true$ and $u \in Ext\ (s)$.
- G: from S in $P(\Omega)$ such that: $G(s) = Ext\ (s)$.

A « complete symbolic object » s is such that F(G(s)) = s. Such objects can be selected from the Galois lattice but also, from a partitioning, a hierarchical or a pyramidal clustering, from the most influential individuals in a factorial axis, from a decision tree, etc.

**Example:**

The symbolic data table is given in figure 3. With the choice T ≡ Max and R ≡ ≤ , the Galois lattice obtained from this table is given in the same figure 3. This lattice can be build for instance, by using a generalisation to symbolic data of Chein (1969) or Ganter algorithm (1984) which can be found in Diday ( 1998, 2000) and Pollaillon (1998). The set of all the complete symbolic objects and their extent which are the vertices of this lattice are the following:

$s_1 : a_1(w) = [y_1(w) \subseteq O_1] \wedge [y_2(w) \subseteq O_2] \wedge [y_3(w) \subseteq O_3]$, $Ext(s_1) = \{w_1, w_2, w_3, w_4\}$

$s_2 : a_2(w) = [y_2(w) \subseteq \{e\}] \wedge [y_3(w) \subseteq \{g,h\}]$, $Ext(s_2) = \{w_1, w_2, w_4\}$

$s_3 : a_3(w) = [y_1(w) \subseteq \{c\}]$, $Ext(s_3) = \{w_2, w_3\}$

$s_4 : a_4(w) = [y_1(w) \subseteq \{a,b\}] \wedge [y_2(w) = \varnothing] \wedge [y_3(w) \subseteq \{g,h\}]$, $Ext(s_4) = \{w_1, w_2\}$

$s_5 : a_5(w) = [y_2(w) \subseteq \{e\}] \wedge [y_3(w) \subseteq \{h\}]$, $Ext(s_5) = \{w_4\}$

$s_6 : a_6(w) = [y_1(w) \subseteq \{a,b\}] \wedge [y_2(w) = \varnothing] \wedge [y_3(w) \subseteq \{g\}]$, $Ext(s_6) = \{w_1\}$

$s_7 : a_7(w) = [y_1(w) = \{\varnothing\}] \wedge [y_2(w) = \varnothing] \wedge [y_3(w) \subseteq \{g,h\}]$, $Ext(s_7) = \{w_2\}$

$s_8 : a_8(w) = [y_1(w) = \varnothing] \wedge [y_2(w) = \varnothing] \wedge [y_3(w) = \varnothing]$, $Ext(s_8) = \{\varnothing\}$

| | $y_1$ | $y_2$ | $y_3$ |
|---|---|---|---|
| $w_1$ | {a,b} | ∅ | {g} |
| $w_2$ | ∅ | ∅ | {g,h} |
| $w_3$ | {c} | {e,f} | {g,h,i} |
| $w_4$ | {a,b,c} | {e} | {h} |

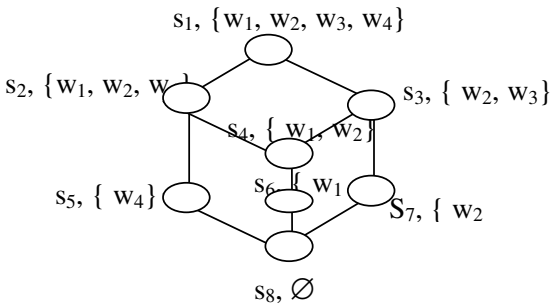



**Figure 3 The conceptual lattice, the symbolic objects associated to the vertices and their extent, obtained from the given Symbolic Data**

**10.2 Mathematical framework of a symbolic data analysis**

Finally we can summarise the mathematical framework of a symbolic data analysis in the following way (figure 4):

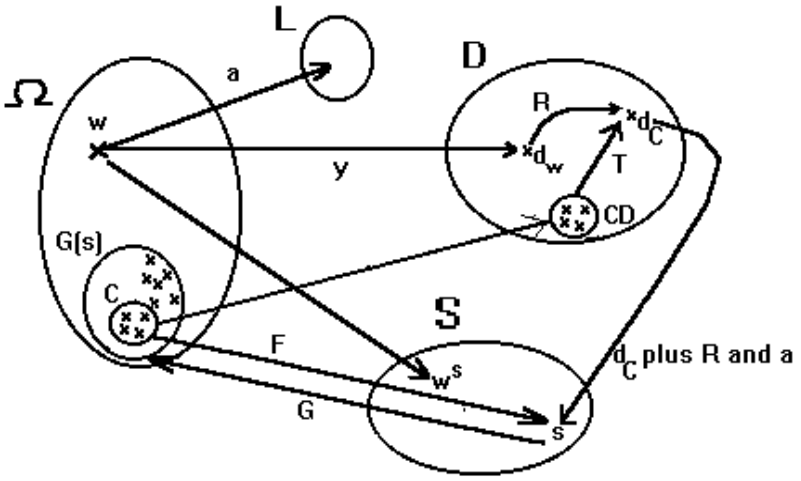



**Figure 4. Ω: set of individuals. D: description set. L= {true, false} or L= [0,1]. S: set of symbolic objects. y: description function. a: membership function from Ω in L= {true, false} or L= [0,1]. R: comparison relation. T: generalisation mapping. F: intension mapping, G: extension mapping. $d_w$ : y(w) = $d_w$ is an individual description. $w^s$ : $w^s$ = F(w) = (a, R, y(w)) is an individual symbolic object. $d_C$ :description of class C. s: intensional symbolic object given by F(C) = (a, R ,$d_C$) where a = [y(w) R $d_C$]. G(s) is the extension of s. Modeling individuals, classes of individuals and concepts**

### 10.3 Statistics of symbolic objects

Instead of studying robustness, reliability and characteristic of symbolic objects, by using their extent, another way consists in using their description part in order, for instance, to find outliers. In that case, we need to extend the notion of "mean", correlation and "standard deviation" to symbolic data, in order to use, for instance, a kind of Fisher and a Student test. In that way, a first effort can be found in Bertrand, Goupil (2000) and Billard, Diday (2000), Rodriguez (2000) where histogram and covariance of interval data are studied.

### 10.4 Quality , robustness and reliability of a symbolic object

It is always possible to choose the operator T in such a way that, the extent of the symbolic object $s_C$ contains the extent of the concept C, but it can contain also individuals which are not in the extent of C. Moreover, for instance in the modal case, we can have some individuals of the extent of C which are not in the extent of $s_C$ in Ω', this depends on the choice of the threshold $\alpha$: Ext $_\alpha$( $s_C$) = {w/ a(w) }$\geq \alpha$}. Therefore we can have two kinds of errors:

i) individuals who satisfy the concept and are not in the extent of $s_C$,

ii) individuals who do not satisfy the concept but are in the extent of $s_C$.

The "quality" and "robustness" of the symbolic object s can then be defined in several ways. We denote $e_1(\alpha)$ and $e_2(\alpha)$ the error percentage in the first and second kind of error. In order to find the best $\alpha$ we can vary $\alpha$ between 0 and 1 and we retain the $\alpha$ which minimises the product $e_1(\alpha)\ e_2(\alpha)$.

In order to validate the symbolic object $s_C$, we suggest the following method. Repeat n times the following steps:

Step1. Obtain a sample $\Omega'$ with replacement from the whole set of given individuals $\Omega$.

Step2. Calculate the symbolic object $s_C$ by following the scheme given in figure 2.

Step3. Calculate the extent of $s_C$ in $\Omega'$.

Step4. Calculate the errors of kind i) and ii).

The quality and robustness of the symbolic object $s_C$ is the higher when the mean and the mean square of the two histograms of the frequency of errors of kind i) and ii) is the lowest. In other words, let X1 (resp. X2) be the random variable which associates to each sample the frequency of error of type i (resp. ii). Then, the lowest is the mean and the mean square of these two random variables, the higher is the quality and robustness of the symbolic object $s_C$.

The "reliability" of the membership of an individual w to the extent of $s_C$ can be measured by the mean m(w) of the $a_C(w)$. More precisely, if the i th sample gives the value $a_i(w)$ then $m(w) = \sum_{i=1,n} a_i(w)/n$ and the reliability of $s_C$ can be defined by :

$W(s_C) = \sum_{w \in Ext(C)} m(w) / \mid Ext(C) \mid$. The higher (i.e. the closest to 1) $W(s_C)$ the better is the reliability of $s_C$. The "sensibility" of $s_C$ can be measured by:

$$W'(s_C) = \sum_{w \in Ext(C)} \sigma(w) / \mid Ext(C) \mid \text{ with } \sigma(w)^2 = \sum_{i=1,n} (a_i(w) - m(w))^2 / n.$$

These measures seem natural but their precise study remains to be done in comparison with other measures of such kind.

## 11. Some symbolic data analysis methods

### 11.1 Main principle

Symbolic Data Analysis methods are mainly characterised by the following principle:

i) they start as input with a symbolic data table and they give as output a set of symbolic objects. These symbolic objects give explanation of the results in a language close to the one of the user and moreover have several advantages mentioned in section 13).

ii) They use efficient generalization processes during the algorithms in order to select the best variables and individuals.

iii) They give graphical descriptions taking account of the internal variation of the symbolic objects.

### 11.2 The main methods

The following methods are developed in Bock, Diday (2000) or in the SODAS (2004) free software. More details are given on some of the modules in the appendix.

#### 11.2.1 Descriptive SDA

- From Database to symbolic description (called DB2SO in SODAS)
- Inter-active and ergonomic graphical representation of symbolic objects (called VIEW in SODAS).
- Extension of elementary descriptive statistics (called DSTAT) to symbolic data (histograms, mean, min, max, biplot, etc. from a symbolic data table).
- Principal Component (called SPCA in SODAS). The output of this methods preserves the internal variation of the input data in the sense that the individuals are not represented in the factorial plane by a point as usual but by a rectangle which allows the definition of a symbolic object with explanatory factorial axes as variables.

#### 11.2.2 Classification of symbolic data

- partitioning, hierarchical or pyramidal clustering (HIPYR).
- Kohonen mapping of a set of individuals described by a symbolic data table such that each class be associated with a complete symbolic object (called SYKSOM) .
- Top-down hierarchical tree (called DIV).
- dissimilarities between Boolean or probabilistic symbolic objects.
-

#### 11.2.3 Predictive SDA

- Discriminate Factorial Analysis of a symbolic data table (called SFDA)
- Extension of decision trees
- Regression on symbolic data
- Generalisation by a disjunction of symbolic objects of a class of individuals described in a standard way.

Explanation of the intial correlation between variables inside each category , extending rule extraction algorithms to symbolic data, symbolic 3-D Clustering.etc.

## 12. Symbolic Data Analysis in the SODAS software

### 12.1 The general aim

The general aim of SODAS can be stated in the following way: building symbolic data in order to describe categories, analyse them by Symbolic Data Analysis. For instance, if a set of households is characterised by its region, the number of bedrooms and of dining-living, its socio-economic group, we obtain a data table of the kind of table 4:

| Household number | Region | Bedroom | Dining-Living | Socio-Econ group |
|---|---|---|---|---|
| 11404 | Northern-Metropolitan | 2 | 1 | 1 |
| 11405 | Northern-Metropolitan | 2 | 1 | 3 |
| 11406 | Northern-Metropolitan | 1 | 3 | 3 |
| 12111 | Northern-Metropolitan | | | |
| 12112 | East anglia | 1 | 3 | 3 |
| 12112 | East anglia | 2 | 2 | 1 |
| 12112 | GreaterLondon N-E | 1 | 2 | 3 |

Table 4 : Standard Data Table where the units are Households

In census data there is a huge set of households. In order to compare the regions, we can summarise them by describing each region by the households of their inhabitants. In order to do so, we delete the first column of this table and we obtain table 5:

| Region | Bedroom | Dining-Liv | Socio-Ec gr |
|---|---|---|---|
| Northern-Metropolitan | 2 | 1 | 1 |
| Northern-Metropolitan | 2 | 1 | 3 |
| Northern-Metropolitan | 1 | 3 | 3 |
| Northern-Metropolitan | | | |
| East-anglia | 1 | 3 | 3 |
| East-anglia | 2 | 2 | 1 |
| East-anglia | 1 | 2 | 3 |
| Greater London North-East | | | |

Table 5: The first column of table 4 concerning the household number has been deleted.

We can now describe each town by the histogram of the categories of each variable. This is done in table 6 which is a symbolic data table as each cell contains a histogram and not a quantitative or categorical number as in standard data tables. It is easy to see that standard data analysis methods will not apply in the same way with these kind of symbolic data. For instance that a decision tree will not be the same if the variables are categories and each cell of the associated data table contains a frequency and if the variables are symbolic and each cell contains a histogram. In the first case each branch of the decision tree represents an interval of frequency (for instance, "the frequency of the category [20, 30] years old is less then 0.3"), whereas in the second case it represents an interval of values ( for instance, "the age is less then 50 years old"). For more details see in Bock, Diday (2000) the chapter 11.

| Region | Bedroom | Dining-Living | Socio-Ec gr |
|---|---|---|---|
| Northern-Metropolitan | (2\3) 2, (1\3) 3 | (2\3) 1, (1\3) 3 | (1\3) 1, (2\3) 3 |
| | | | |
| | | | |
| East-anglia | (2\3) 1, (1\3) 2 | (2\3) 2, (1\3) 3 | (1\3) 1, (2\3) 3 |
| | | | |
| Greater London | | | |

Table 6: A symbolic data table where the units are now the regions

The main steps for a symbolic data analysis in SODAS can then be defined as following:
If there is more than one data table, put the data in a relational database (ORACLE, ACCESS, ...). Then, define a context by giving: the units (individuals, households,...), the classes (regions, socio-economics groups,...), the descriptive variables of the units. Then, build a symbolic data table where the units are the preceding classes, the descriptions of each class is obtained by a histogram as in table 6 or by a generalisation process applied to its members. This is done by a computer program of SODAS called "DB2SO" (from Databases To Symbolic Objects). Finally, apply to this symbolic data table, symbolic data analysis methods (histogram of each symbolic variable, dissimilarities between symbolic descriptions, clustering, factorial analysis, discrimination of a symbolic data table, graphical visualisation of symbolic descriptions, ...).

**12.2 Examples of applications strategy in SODAS**

We start from data provided by the three Statistical institute involved in SODAS (ONS (England), INE (Portugal), EUSTAT (Span)), as household consuming, census, labour force

survey or road transportation. Units are for instance, " unemployment type " defined by each category of a new variable obtained by the cartesian product: "unemployment people categories x age categories x countries" given by a query to the relational database. Then, DB2SO associates to each unit a symbolic description. Hence, we get a symbolic data table on which symbolic data analysis methods can be applied. In order to summarize and to get an overview on this symbolic data table, we can for instance, apply the following steps: we apply DIV (see chapter 11 in Bock, Diday (2000) ) which provides classes of units. It is then possible to apply again DB2SO on the same units but with the classes given by DIV. Therefore, each class represents a set of regions or a set of unemployment type. Hence, we obtain a new symbolic data table where each unit represents one of these classes. Several symbolic data analysis methods can then be applied: for instance, a principal component analysis (PCA, see chapter 9 in Bock, Diday (2000)) in order to get a graphical overview on these classes, a graphical visualisation of each class by "stars" (see chapter 7 in Bock, Diday (2000)), a description of each class by a disjunction of assertions (DSD, see section 9.4 in Bock, Diday (2000)), etc.

### 12.3 SODAS software overview

In figure 5 an overview on the SODAS software is given. The input of DB2SO (see chapter 5 in Bock, Diday (2000)) is a query to a database. Its output is a symbolic data table. Having obtained this data table any symbolic data analysis method can be applied.

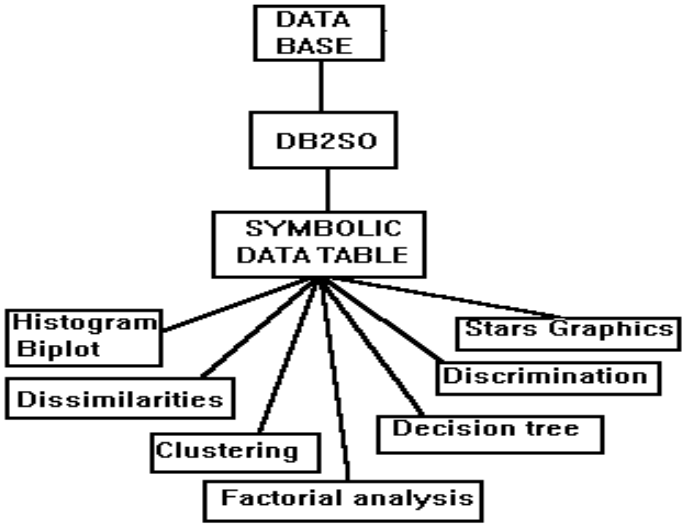



**Figure 5: A SODAS software overview**

## 13 Some advantages in the use of symbolic objects and concept propagation

### 13.1 Some advantages in the use of symbolic objects

We can observe at least six kinds of advantages in the use of symbolic objects.

. They give a summary of the original symbolic data table in an explanatory way, (i.e. close to the initial language of the user) by expressing descriptions based on properties concerning the initial variables or meaningful variables (such as indicators obtained by regression or factorial axes).

. They can be easily transformed in terms of a query of a Database and so they can be used in order to propagate concepts between databases (for instance, from one country to another country).

. By being independent of the initial data table they are able to identify any matching individual described in any data table.

. In the use of their descriptive part, they are able to give a new symbolic data table of higher level on which a symbolic data analysis of second level can be applied.

. In order to characterise a concept, they are able to join easily several properties based on different variables coming from different relations in a Database and different samples of a population.

. In order to apply exploratory data analysis to several databases, instead of merging them in a huge database, an alternative is to summarise each Database by symbolic objects and then to apply Symbolic Data Analysis to the whole set of obtained symbolic objects.

### 13.2 Concept propagation

This propagation is interesting for instance, in order to see if concepts obtained on a survey in a contry remains in a survey done in another country or at another time. In order to propagate concepts from a database to another database there is a module called SO2DB (i.e. Symbolic Objects to Database) due to Donato and al.. This prpagation is based on the following steps. First to extract symbolic objects from the clustering, factorial analysis, decision tree or discrimination (standard or symbolic) methods. Second, to induce from these symbolic objects, a new symbolic data table in order to study them, by a symbolic data analysis of higher level. Third, to select the "best" symbolic objects and prototypes, by using good criteria . Fourth, to propagate the obtained symbolic objects (the concepts that they represent). This propagation can be done towards the same Database for instance, at different times (in order to study the time evolution of the retained concepts) or towards other databases

associated with different countries. In any case, we have to compare sets of concepts and their associated symbolic objects obtained from different databases. This may be done in several ways. For instance, by looking for a consensus tree or pyramid, between the concepts obtained in two different countries. Among many other possibilities, we can also calculate the extent of the symbolic objects obtained from a country in another country and then comparing the concepts associated to the symbolic objects of the first country to the concepts of the second country induced by the "complete symbolic objects" obtained from these extension. An overview on the next steps for the research and development of SODAS project are given in figure 6.

**Conclusion**

The need to extend standard data analysis methods (exploratory, clustering, factorial analysis, discrimination, ...) to symbolic data tables in order to extract new knowledge, is increasing due to the expansion of information technology, now able to store an increasing amount of huge data sets. This need, has led to a new methodology called "Symbolic Data Analysis" whose aim is to extend standard data analysis methods (exploratory, clustering, factorial analysis, discrimination, decision trees, etc.) to new kind of data table called "symbolic data table" and to give more explanatory results expressed by real world concepts mathematically represented by easily readable "symbolic objects". The aim of the EUROSTAT European Community project called SODAS for a « Symbolic Official Data Analysis System» in which 17 institutions of 9 European countries are concerned was to produce a first software of Symbolic Data Analysis. Three Official Statistical Institutions were involved in this project: EUSTAT (Span), INE (Portugal) and ONS (England). An example of future applications proposed on their Census data consists in finding clusters of unemployed people and their associated mined symbolic objects in a country, calculating its extent in the census of another country and describing this extent by new symbolic objects in order to compare the behavior of the two countries. To this end, several new theoretical developments are needed such as the selection and the stochastic convergence of symbolic objects. Also needed, are developments permitting one to obtain the consensus between set of symbolic objects and their associated concepts extracted from different databases. New software developments are also needed as a tool in order to be able to transform a symbolic object extracted from a database in a query of this database or of another database. This new tool may be called SO2DB as it is complementary to the actual DB2SO. Moreover, the next steps will be to improve the actual SDA methods (robustness, validity of the results, extending standard tests to symbolic data,

etc.) and extend the symbolic data analysis methodology to regression, multidimensional scaling, neural networks etc.

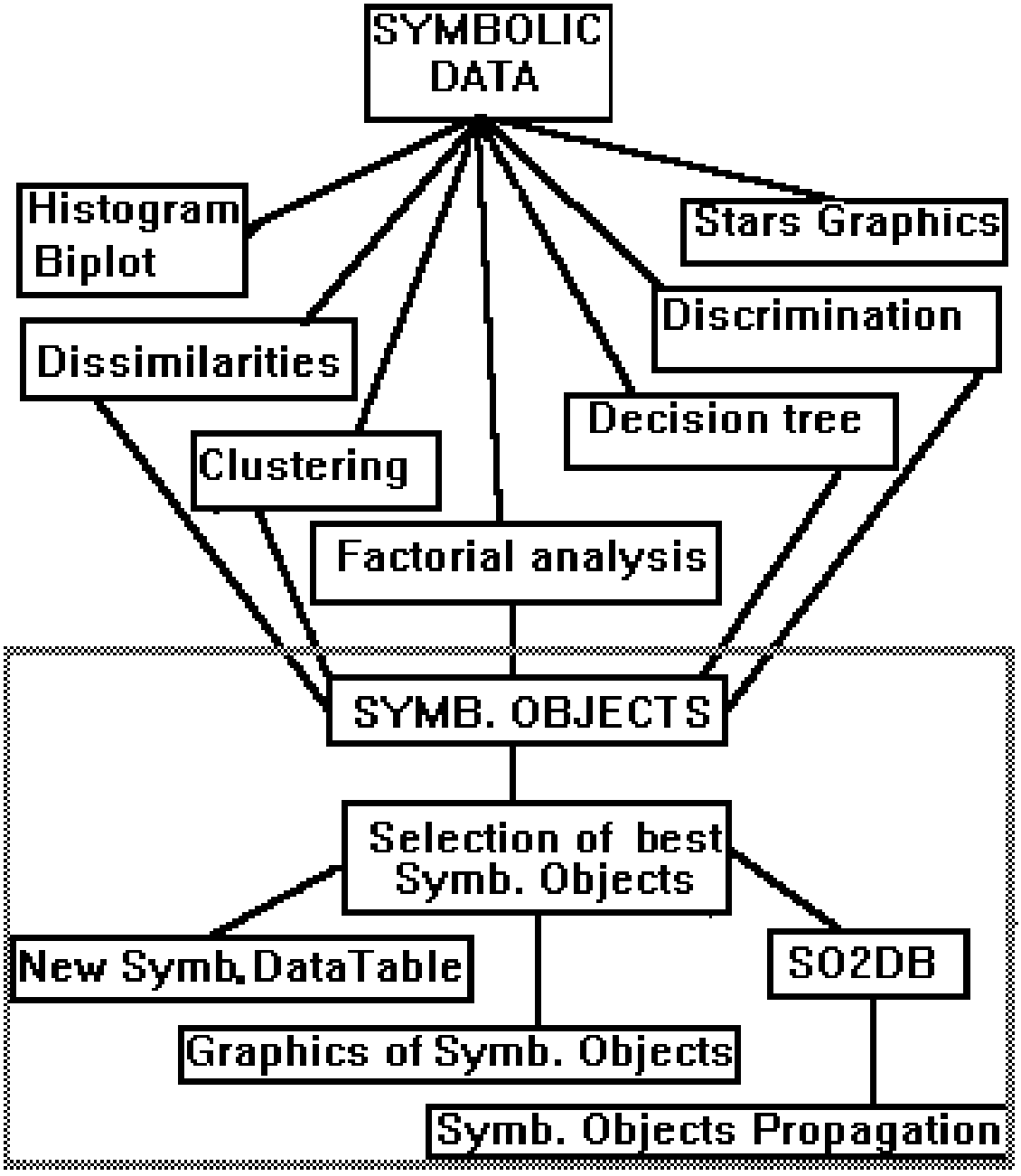



**Figure 6: Symbolic objects propagation.**

Concerning the free SODAS software much information can be found in Bock, Diday (2000) and in the synthesis paper: L. Billard, E. Diday "From the statistic of Data to the statistic of Knowledge". JASA. Journal of the American Statistical Association. June(2003)..

In the Web site **www.ceremade.dauphine.fr/%7Etouati/sodas-pagegarde.htm** you can load SODAS on your PC computer and get much informations on the methods and many examples of applications in numerous domains. Recent papers can also be found in the Electronical Journal of SDA (JSDA) at **director@unina2.it**.

**References**


- Y.J.P. Aboa(2001) "Méthodes de segmentation sur un tableau de variables aléatoires". Thèse de Doctorat. University Paris IX Dauphine.
- Adanson M. (1757) "Histoire Naturelle du Sénégal- Coquillages". Bauche Paris.
- Aristotle (IV BC) "Organon" Vol. I Catégories, II De l'interprétation. J. Vrin edit. (Paris) (1994).
- Arnault A., Nicole P. (1662), "La logique ou l'art de penser", Froman, Stutgart (1965).
- Auriol E. (1995) "Intégration d'approches symboliques pour le raisonnement à partir d'exemples" Thèse de doctorat, Université Paris 9 Dauphine.
- Barbut M., Monjardet B. (1971°, "Ordre et classification", T.2 Hachette, Paris.
- Belson (1959), "Matching and prediction on the principle of biological classification", Applied Statistics, vol. VIII.
- Benzecri J.P. et al. (1973) "L'Analyse de Données", Dunod, Paris.
- Bertrand P. , Goupil F. (2000) "Descriptive statistics for Symbolic Data". In "Analysis of Symbolic Data". Bock, Diday edit. Study in Classification, Data Analysis and Knowledge Organisation. Springer Verlag.
- Bertrand P. (1986) "Etude de la représentation pyramidale", Thèse de 3° cycle, Université Paris IX-Dauphine.
- Billard L., Diday E. (2000) "Regression Analysis for Interval-Valued Data" . In Proc. Of IFCS-2000. "Data Analysis, Classification and related methods". Kiers and all editors. Springer Verlag.
- L. Billard, E. Diday (2003) "From the statistic of Data to the statistic of Knowledge". JASA. Journal of the American Statistical Association. June.
- Bock H.H. (1974) "Automatische Klassifikation". Vandenhoeck and Ruprecht, Gottingen.
- Bock H.H., Diday E. (2000) "Analysis of Symbolic Data". Study in Classification, Data Analysis and Knowledge Organisation. Springer Verlag.
- Breiman L., Friedman J.H., Olsken R.A., Stone C.S. (1984) "Classification and regression trees", Belmont, Wadsworth.
- Brito P. (1991) "Analyse des données symboliques. Pyramides d'héritage". Thèse de doctorat, Université Paris 9 Dauphine.
- Brito P., Diday E. (1991) "Pyramidal representation of symbolic objects" NATO ASI Series, Vol. F 61. Proc. Knowledge Data and computer-assisted Decisions. Schader and Gaul edit. Springer-Verlag.
- Brito P. (1994) "Order structure of symbolic assertion objects". IEEE TR. on Knowledge and Data Engineering Vol.6, n° 5, October.
- Bandemer H., Nather W. (1992) "Fuzzy Data Analysis". Kluwer Academic Publisher.

- Cazes P., Chouakria A., Diday E., Schecktman Y.(1997)) "Extension de l'Analyse en Composantes Principales à des données intervalles". Revue de Statistiques Appliquée, vol. XXXVIII, n°3, 1990,pp 35-51.
- Celeux G., Diday E., Govaert G., Lechevallier Y., Ralambondrainy H. (1989), "Classification Automatique: environnement Statistique et Informatique". Dunod.
- Changeux J.P. (1983) "L'homme neuronal". Fayard, Collection Pluriel.
- Chavent M. (1997) "Analyse des Données symboliques. Une méthode divisive de classification". Thèse de doctorat, Université Paris 9 Dauphine.
- Ciampi A., Diday E., Lebbe J., Périnel E., Vigne (1995) R. " Recursive partition with probabilistically imprecise data". OSDA'95. Editors: Diday, Lechevallier, Opitz Springer Verlag (1996).
- Ciampi, E. Diday, J. Lebbe, E. Perinel, R. Vignes (2000) "Growing a tree classifier with imprecise data" . Pattern Recognition letters 21, pp 787-803.
- Conruyt N. (1994) "Amélioration de la robustesse des systèmes d'aide à la description, à la classification et à la détermination des objets biologiques. Thèse de doctorat, Université Paris 9 Dauphine.
- De Carvalho F.A.T. (1998) a a "New metrics for constrained boolean symbolic objects" Proc. KESDA'98, Eurostat. Luxembourg.
- De Carvalho F.A.T. (1998) b "Statistical proximity functions of boolean symbolic objects based on histograms" IFCS, Roma, Springer-Verlag.
- Diday E.(1971) "La méthode des nuées dynamiques" ; Revue de Statist. Appliquée. Vol XIX, n° 2, pp. 19-34.
- Diday E.(1976) "Sélection typologique de variables". Rapport INRIA. Rocquencourt 78150, France.
- Diday E.(1976) "Cluster analysis" in K.S. Fu (ed.). Digital Pattern Recognition. Springer Verlag. PP. 47-94.
- Diday E. et al. (1979) "Optimisation en classification automatique". INRIA edit. Rocquencourt 78150, France.
- Diday E., Govaert G., Lechevallier Y., Sidi J. (1980) "Clustering in Pattern Recognition". Proceed. NATO Adv. Study Institute on Digital Processing and Analysis, Bonas, J.C. Simon edit.
- Diday E. (1984) "Une représentation visuelle des classes empiétantes". Rapport INRIA n° 291. Rocquencourt 78150, France.
- Diday E., Lemaire J., Pouget J., Testu F. (1984) "Eléments d'Analyse des données". Dunod, Paris.
- Diday E. (1986) "Orders and overlapping clusters by pyramids". Proceed. Multidimensional Data Analysis. Edits. J.D. De Loeuw et al, DSWO Press, Leiden, The Netherlands.
- Diday E. (1987 a) "The symbolic aproach in clustering and related methods of Data Analysis" in "Classification and Related Methods of Data Analysis", Proc. IFCS, Aachen, Germany. H. Bock ed.North-Holland.
- Diday E. (1987 b) "Introduction à l'approche symbolique en Analyse des Données ". Première Journées Symbolique-Numérique. Université Paris IX Dauphine. Décembre 1987.
- Diday E. (1989) "Introduction à l'approche symbolique en analyse des données". RAIRO (Revue, d'Automatique, d'informatique et de Recherche Opérationnelle), vol. 23, n°2.
- Diday E. (1991) "Des objets de l'Analyse des Données à ceux de l'Analyse des Connaissances" in "Induction symbolique et numérique". Y. Kodratoff and E. Diday edit. CEPADUES-EDITIONS, Toulouse, France.

- Diday E. (1995) " Probabilist, possibilist and belief objects for knowledge analysis " .Annals of Operations Research . 55, 227-276.
- Diday E., Emilion R. (1995) "Lattices and Capacities in Analysis of Probabilist Objects". Proceed. of OSDA'95 (Ordinal and Symbolic Data Analysis). Springer Verlag Editor (1996).
- Diday E., Emilion R. (1997) " Treillis de Galois maximaux et Capacités de Choquet" Compte rendu à l'Académie des Sciences. Analyse Mathématique, t. 324, série 1.
- Diday E., Emilion R., Hillali Y. (1996) "Symbolic data analysis of probabilist objects by capacities and credibilities. XXXVIII Societa Italiana Di Statistica. Rimini, Italy.
- Diday E.(1998) "L'Analyse des Données Symboliques: un cadre théorique et des outils" . Cahiers du CEREMADE n° 9821.
- E. Diday (2000) "Analyse des données symboliques: théorie et outil pour la fouille de connaissances" TSI (Technique et Science Informatiques). Vol 19, n°1-2-3 , Janvier 2000.
- Diday E. (2000) "Partitioning concepts described by distributions with copulas modeling" OSDA '2000. Bruxelles.
- E. Diday, O. Rodriguez, P. Brito (2000) "Algoritmos para la classification pyramidal simbolica" Teoría y Aplicaciones, Vol. 7, N. 1-2, 2001, pp. 23-42.
- E. Diday, Y. Kodratoff, P. Brito, M. Moulet (2000): "Induction symbolique numérique à partir de données". Cépadues. 31100 Toulouse. www.editions-cepadues.fr. 442 pages.
- E. Diday (2002) "Mixture Distributions of Distributions by Copulas". Proceedings IFCS'2002, In Krzysztof Jajuga et al (Eds.): Data Analysis, Classification and Clustering Methods Heidelberg, Springer-Verlag Springer Verlag.
- Diday E., Emilion R. (2003) "Maximal stochastic Galois Lattice". DAM. Journal of Discrete Applied Mathematics. Volume 127, Issue 2 , 15 April 2003, Pages 271-284.
- Esposito F., Malerba D., Lisi F. (1998) "Flexible matching of Boolean symbolic objects" Proc. NTTS'98 Sorrento, Italy. Nanopoulos, Garonna, Lauro edit. Eurostat (Luxembourg).
- Ferraris, Gettler-Summa, C. Pardoux, H. Tong (1995) "Knowlege extraction using stochastic matrices: Application to elaborate a fishing strategy" Proc. Ordinal and Symbolic Data Analysis. Paris ; Diday, Lechevallier, Opitz edit. Springer Studies in Classification.
- Fisher D.H., Langley P. (1986) "Conceptual clustering and its relation to Numerical Taxonomy". Workshop on Artificial Intelligence and Statistics" Addisson-Wesley, W. Gale édit.
- Fisher D.H.,(1987) a, "Conceptual clustering learning from examples and inference". Proceed. 4th Workshop on Machine Learning. Irvine, California.
- Ganascia J.G. (1991) "Charade: apprentissage de bases de connaissances". Cepadues, Kodratoff, Diday edit.
- Gettler-Summa M. (1992) "Factorial axis interpretation by symbolic objects". Journées - Symbolique - Numérique. Université Paris IX- Dauphine. Lise-Ceremade.
- Gettler-Summa M. (1997) "Symbolic marking: application on car accidents scenari" Proc. AMSDA, Capri, Italy.
- Gigout E. (1998) " Graphical interpretation of symbolic objects resulting from data mining". Proc. KESDA'98, Eurostat. Luxembourg.
- Gowda K.C., Diday E. (1992) "Symbolic clustering using a new similarity measure". IEEE Trans. Syst. Man and Cybernet. 22 (2), 368-378.
- Gower J.C. (1974) "Maximal predictive classification". Biomet. Vol. 30, p. 643-644.
- Hayes-Roth F., McDermott J. (1978) "An interference matching technique for inducing abstractions"Comm. ACM. Artificial Intelligence, Language processing.

- Hebrail G. (1996) " SODAS (Symbolic Official Data Analysis System) ". Proceedings of IFCS'96, Kobe , Japan. Springer Verlag.
- Jambu M. (1978) "Classification Automatique pour l'Analyse des Données". Dunod, Paris.
- Jardine N., Sibson R. (1971) "Mathematical Taxonomy". John-Wiley and Sons. New-York.
- Jussieu A.L. (1748) "Taxonomy. Coup d'oeil sur l'histoire et les principes des classifications botaniques". Dictionnaire d'Histoire Universelle.
- Lance G.N. , Williams W.T. (1967) "A general theory of Classification sorting strategies: hierarchical systems". Comp. Jorn. Vol. 9 n°4.
- Langley P., Sage S. (1984) "Conceptual clustering as discrimination learning". Proceed. Fifth Biennial Conf. the Canadian Soc. for Comp. Studies of Intelligence.
- Labowitz M. (1983) " Generalization from natural language text" Cognit. Science 7, 1.
- Lauro C., Palumbo F. (1998) "New approaches to Principal Component Analysis of Interval Data". Proc. NTTS'98 Sorrento, Italy. Nanopoulos, Garonna, Lauro edit. Eurostat (Luxembourg).
- Lebart L., Morineau A., Piron M. (1995) "Statistique Exploratoire Multidimensionnelle" . Dunod, Paris.
- Lebbe J. and Vignes R. (1991) "Génération de graphes d'identification à partir de descriptions de concepts", in Induction Symbolique-Numérique. Kodratoff, Diday edit. Cepadues (Toulouse).
- Lerman I.C. (1970) "Les bases de la classification automatique" Gautier-Villars Paris.
- Noirhomme-Fraiture, Rouard M. (1998) " Representation of Sub-Populations and Correlation with Zoom Star". Proc. NTTS'98 Sorrento, Italy. Nanopoulos, Garonna, Lauro edit. Eurostat (Luxembourg).
- Mfoumoune E. (1998) "Les aspects algorithmiques de la classification ascendante pyramidale et incrémentale" . Thèse de doctorat, Université Paris 9 Dauphine.
- Michalski, R. (1973), Aqual/1 -Computer Implementation of a variable-valued logic system VL1 and examples in Pattern Recognition". Proc. Int. Joint Conf. on Pattern Recognition, Washington D.C., pp 3-17.
- Michalski R., Step R.E. (1983) "Automated construction of classifications Conceptual Clustering versus Numerical Taxonomy", IEEE Trans. on Pattern Analysis and Machine Intelligence. Vol. 5, n°4.
- Michalski R., Diday E., Step R.E. (1982) " A recent advances in Data Analysis: clustering objects into classes characterized by conjonctive concepts". Progress in Pattern Recognition , vol 1. L; Kanal and A. Rosenfeld Eds.
- Morgan J.N., Sonquist J.A. (1963) "Problems in the analysis of survey data : a proposal". J.A.S.A. 58, p. 417-434.
- Pankhurst R.J. (1978) "Biological identification. The principle and practice of identificatin methods in biology". London, Edward Arnold.
- Payne R.W. (1975) "Genkey: a program for construction diagnostic keys". Biological Identification with Computer .Pankhurst edit. P. 65-72. Acad. Press. London
- Périnel E. (1996) "Segmentation et Analyse de Données Symboliques: Application à des données Probabilistes Imprécises". Thèse de doctorat, Université Paris 9 Dauphine.
- Pollaillon G., Diday E. (1997) " Galois lattices of symbolic objects " Rapport du Ceremade University Paris9- Dauphine (February).
- Pollaillon G. (1998) "Organisation et interprétation par les treillis de Galois de données de type multivalué, intervalle ou histogramme". Thèse de doctorat, Université Paris 9 Dauphine.

- Rasson J.P., Lissoir S. (1998) "Symbolic Kernel discriminante analysis" Proc. NTTS'98 Sorrento, Italy. Nanopoulos, Garonna, Lauro edit. Eurostat (Luxembourg).
- Rodriguez O. (2000) "Classification et modèles linéaires en Analyse des Données Symboliques". Thèse de doctorat, University Paris 9 Dauphine.
- Quinlan J.R. (1986) "Induction of decision trees". Machine Learning 1, pp 81-106. Kluwer Acad. Publishers, Boston.
- Ralambondrainy H. (1991) "Apprentissage dans le contexte d'un schéma de base de Données" Kodratoff, Diday edit. CEPADUES, Toulouse.
- O. Rodriguez (2000) "Classification et Modèles Linéaires en Analyse des Données Symboliques". Thèse de Doctorat. University Paris IX Dauphine.
- Rosch E. (1978) "Principle of categorization". E. Rosch, B. Lloyd edits. Cognition and Categorization , pp 27-48 . Hillsdale, N.J.: Erlbaum.
- Roux M. (1985) "Algorithmes de classification", Masson..
- Saporta G. (1990) "Probabilités, Analyse des Données et Statistiques". Edit. Technip Paris.
- Schweizer B. (1985) "Distributions are the numbers of the futur". Proc. sec. Napoli Meeting on "The mathematics of fuzzy systems". Instituto di Mathematica delle Faculta di Mathematica delle Faculta di Achitectura, Universita degli studi di Napoli. p. 137-149.
- Schweizer B., Sklar A. (1983) " Probabilist metric spaces ". Elsever North-Holland, New-York.
- Sneath P.H.A., Sokal R.R. (1973) "Numerical Taxonomy" Freeman and Comp. Publishers. San Francisco.
- Sowa J. (1984) Conceptual Structures: Information processing in mind and machine. Addison Wesley.
- Stéphan V. (1998) "Construction d'objets symboliques par synthèse des résultats de requêtes SQL. Thèse de doctorat, Université Paris 9 Dauphine.
- Tukey J. W. (1958) "Exploratory Data Analysis". Addisson Wesley, Reading, Mass.
- Vignes (1991) "Caractérisation automatique de groupes biologiques". Thèse de doctorat, Université Paris 9 Dauphine.
- Verde R., F.A.T. De Carvalho (1998) "Dependance rules influence on factorial representation of boolean symbolic objects". Proc. KESDA'98, Eurostat. Luxembourg.
- Vrac M. (2002) Analyse et Modélisation de Données probabilistes par Décomposition de Mélange de Copules et Application à la Climatologie.
- Vrac M., Diday E., Winsberg S., Limam M.M. (2002) "Symbolic class description" Proceedings IFCS'2002, In Krzysztof Jajuga et al (Eds.): Data Analysis, Classification and Clustering Methods Heidelberg, Springer-Verlag Springer Verlag.
- Wagner H. (1973) "Begriff", Hanbuck Philosophischer Grundbegriffe, eds H. Krungs, H.M. Baumgartner and C. Wild, Kosel, Munchen ; PP. 191- 209.
- Ward J.H. (1963) 3hierarchical groupings to optimize an objective function". J. Amer. Stat. Assoc. 58, pp. 236-244.
- Wille R. (1982) "Restructuring lattice theory: an approach based on hierarchies of concepts." Proceed. Symp. Ordered Sets (I. Rival ed.), Reidel, Dordrecht-Boston.
- Wille R. (1989) "Knowledge Acquisition by methods of formal concepts analysis, in Data Analysis, Learning symbolic and Numeric Knowledge. Diday edit. Nova Science Publishers.
- Winston P. (1979) "Artificial Intelligence". Addison Wesley.
- Ziani D. (1996) "Sélection de variables sur un ensemble d'objets symboliques" Thèse de doctorat, Université Paris 9 Dauphine.

## APPENDIX

**DB2SO** (G. Hébrail, Y. Lechevallier, V. Stephan)

In the DB2SO module, the user can create a set of descriptions modeling categories from data stored in a relational database. Two main steps are performed:

a) building of symbolic descriptions corresponding to descriptions of groups of individuals of the initial database (by a generalization process),

b) Refining the description to make them simpler by removing some untypical individuals.

c) Symbolic description, built by the basic process, can be enriched by metadata, still picking up information from the database: DB2SO supports addition of taxonomies in variable domains, mother/daughter variables and rules.

Finally we present some additional features included in DB2SO, which facilitate the user's life in the process of constructing of a set of symbolic objects from a database.

An example is presented as a didactic introduction to construction of symbolic objects from the contents of relational databases.

An interesting property of relational databases is that operations can be defined on tables that produce new tables. Queries on a relational database are defined by a combination of these operations: the result of a query is consequently also structured as a table. The standard language for querying relational databases is the SQL language.

From a query to the initial database which produce a data table (where the first column express the single name of each individual, the second column the name of the category of each individual, the next columns are associated to the descriptive variables of the individuals), DB2SO create three types of symbolic variables:

- The interval of observed values on individuals in the group for numerical variables,
- The list of observed values on individuals in the group for nominal variables,
- The frequency distribution of observed values on individuals in the groups for nominal variables.

The output of DB2SO is an assertion (see section 8.1). This assertion can be refined in order to avoid a too large covering of the category. The idea of the refinement process is to remove from each group some untypical individuals. Each group of individuals is processed separately. With each SO can be associated a function which says if an individual is recognized by a SO or not. In the case of interval variables or Boolean multi-valued ones, this

function says '*yes*' if individual values for every variable are in the interval or list of values of the corresponding OS variables. Untypical individuals are removed and simpler SO are built again from the resting individuals. The following optimization constraints guides the removal of individuals:

•A minimum threshold is defined on the number of individuals of the group still recognized by the refined SO (this threshold is given by the user, typically 80%),

•A volume measure is associated with each SO, representing the amount of subspace its variable values occupy in the Cartesian product of variable domains. The choice of untypical individuals to be removed is done to maximize decrease of volume associated with the refinement process.

Details of this refinement process and a description of the algorithm can be found in Stéphan (1998) and Bock, Diday (2000).

**DSTAT** (P. Bertrand, F. Goupil)

The goal of DSTAT is to extend standard descriptive statistics to the case of symbolic data. It produces:

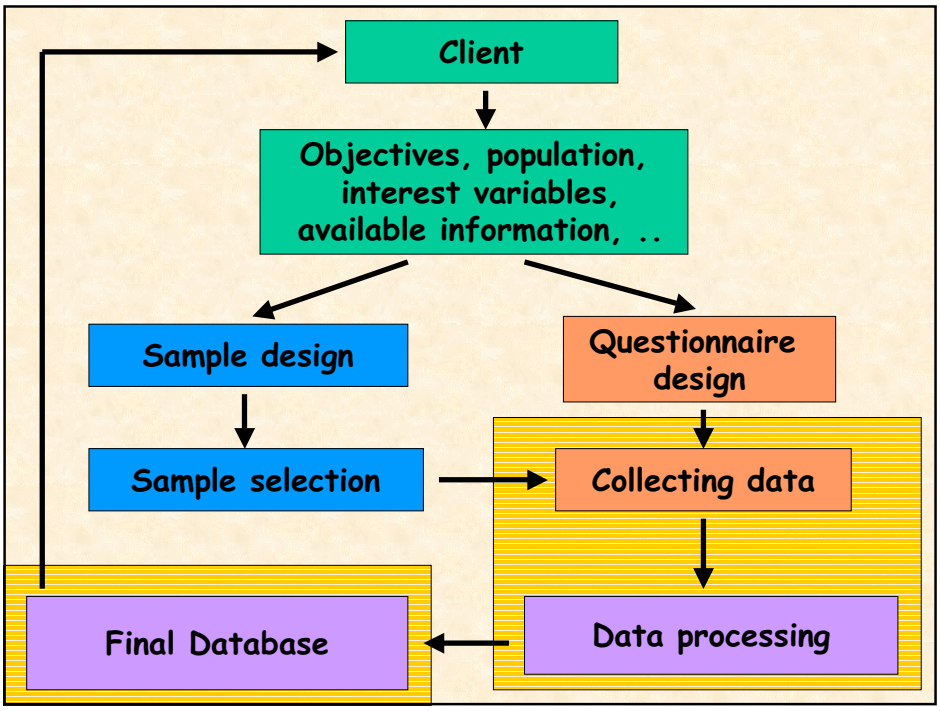
- relative frequencies of each category for variables (called interval variables) whose value for each unit are a unique or several categories.
- relative frequencies for variables whose value for each unit are intervals
- capacities and min/max/mean histogram for variables whose values are frequency distributions.
- biplot for interval variables .

INSTITUTO NACIONAL DE ESTATÍSTICA
PORTUGAL

# Application of Symbolic Data Analysis to Official Statistical Data

Carlos Marcelo
*carlos.marcelo@ine.pt*

T2 - Tutorial on Symbolic Data Analysis
Pisa, 20th September, 2004



## Applications of SDA

- Exploitation of final databases
- Quality Control Process in Statistical Production

## Time Use Survey

*Case Study 1*

## DATA

- **Observed Data**

  Portuguese Time Use Survey 1999

  7021 adults and 34 variables (+ weight)

- **Processing data**

  **BLAISE → ASCII → ACCESS → SODAS**

- **Symbolic Objects**

  - **Sex, Age, daytype (2*5*2=20 SO)**

## VARIABLES

- socio-demographic:
  - region
  - sex
  - age group
  - marital status
  - education level
- activity:
  - occupation
  - economic activity
  - professional status
  - number of hours worked group
- type of day (weekend; workday)

## Variables of perception of time

- **hurried** - Do you feel like you are in a hurry?
  - yes, no
- **hurried_often** - How often do you feel in a hurry?
  - every day, often or sometimes
- **hurried_5years** - Compared to 5 years ago?
  - more hurried, no change, less hurried
- **hurried_3years** - Compared to 3 years ago?
  - more hurried, no change, less hurried
- **hurried_1year** - Compared to last year?
  - more hurried, no change, less hurried

## Variables of perception of time

- **reduce** - do you intend to reduce your rhythm of life next year?
  - yes, no
- **work_time** - Work - Time to accomplish all the tasks?
  - everyday, often; sometimes; rarely; never
- **private_life_time** - Time to accomplish all the tasks?
  - everyday, often; sometimes; rarely; never
- **weekend_time** - Time to accomplish everything that you wanted?
  - everyday, often; sometimes; rarely; never
- **available** - Available time without knowing what to do with it?
  - everyday, often; sometimes; rarely; never

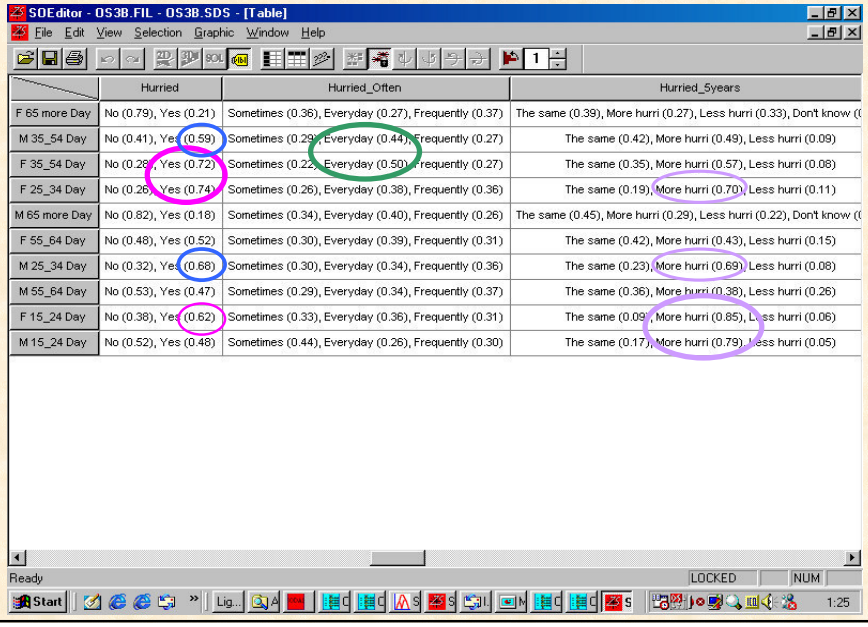
## Example of rule design

**Hurried_Often** is applicable if **Hurried** in {"Yes"}
**Hurried_5years** is applicable if **Hurried** in {"Yes"}
**Hurried_3years** is applicable if **Hurried** in {"Yes"}
**Hurried_1years** is applicable if **Hurried** in {"Yes"}
**Reduce** is applicable
if **Hurried** in {"Yes"}
**Work_Time** is applicable
if **Hurried** in {"Yes"}
**Private_Life_Time** is applicable
if **Hurried** in {"Yes"}
**Weekend_Time** is applicable
if **Hurried** in {"Yes"}
**Available_time** is applicable
if **Hurried** in {"Yes"}


## Variables of duration of activities

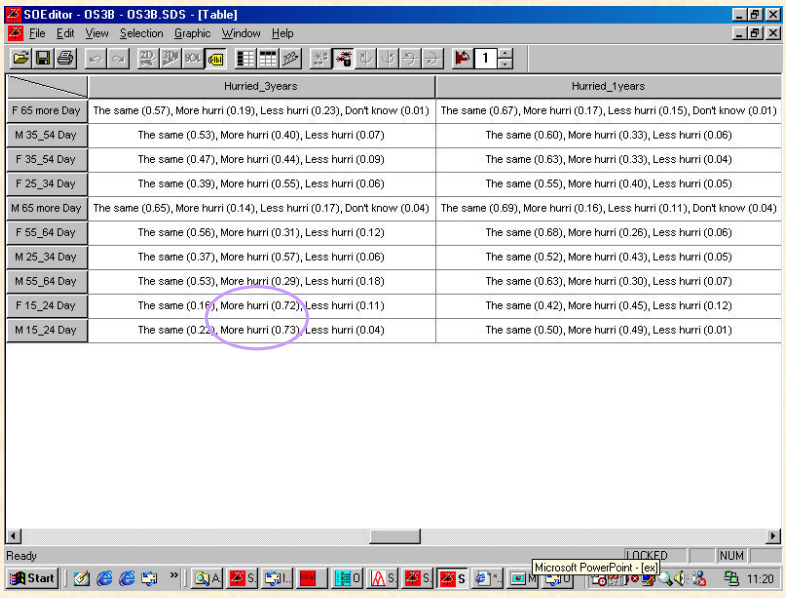
- sleep
- eat
- personal_hig
- profissional_act
- study
- prep_food
- clothes_care
- purchases
- children cares
- social life
- sports
- games
- read
- television_video

# SOE - Perception of time

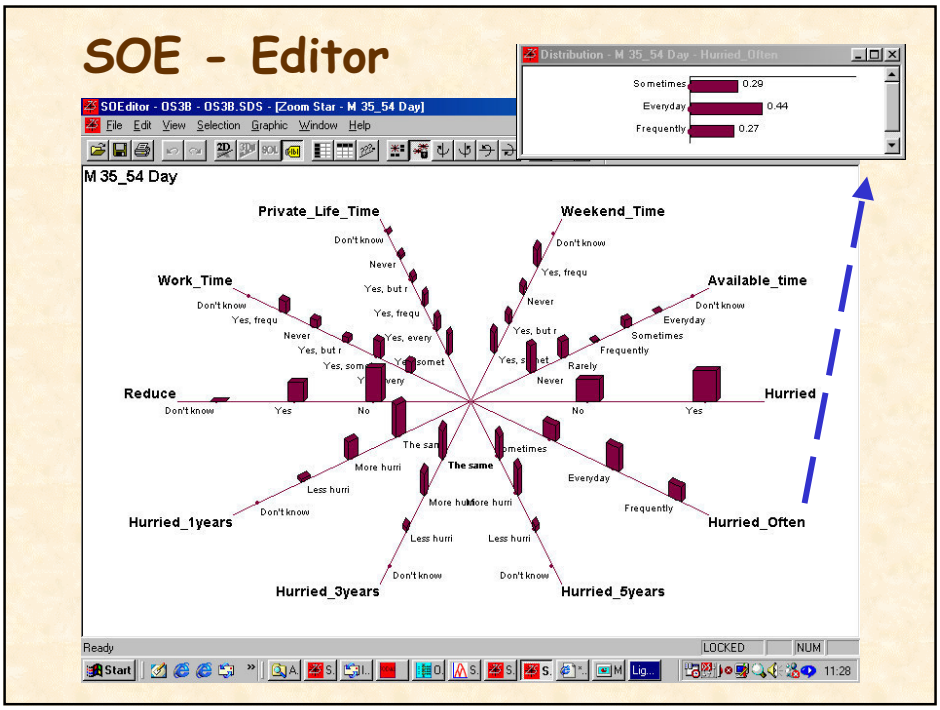


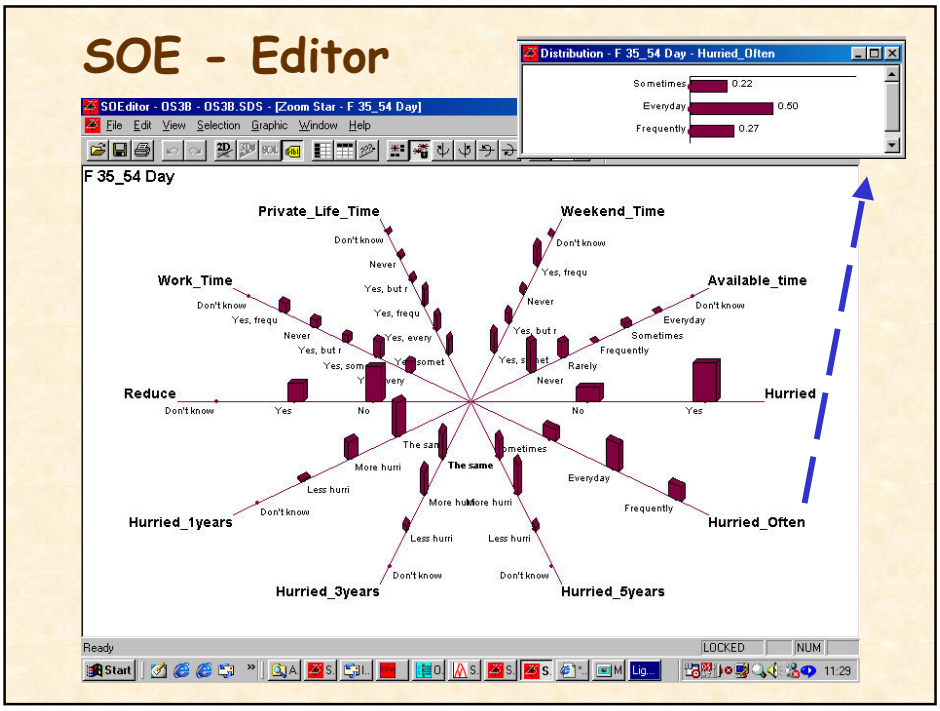

| | Hurried | Hurried_Often | Hurried_5years |
|---|---|---|---|
| F 65 more Day | No (0.79), Yes (0.21) | Sometimes (0.36), Everyday (0.27), Frequently (0.37) | The same (0.39), More hurri (0.27), Less hurri (0.33), Don't know ( |
| M 35_54 Day | No (0.41), Yes (0.59) | Sometimes (0.29), Everyday (0.44), Frequently (0.27) | The same (0.42), More hurri (0.49), Less hurri (0.09) |
| F 35_54 Day | No (0.28), Yes (0.72) | Sometimes (0.22), Everyday (0.50), Frequently (0.27) | The same (0.35), More hurri (0.57), Less hurri (0.08) |
| F 25_34 Day | No (0.26), Yes (0.74) | Sometimes (0.26), Everyday (0.38), Frequently (0.36) | The same (0.19), More hurri (0.70), Less hurri (0.11) |
| M 65 more Day | No (0.82), Yes (0.18) | Sometimes (0.34), Everyday (0.40), Frequently (0.26) | The same (0.45), More hurri (0.29), Less hurri (0.22), Don't know ( |
| F 55_64 Day | No (0.48), Yes (0.52) | Sometimes (0.30), Everyday (0.39), Frequently (0.31) | The same (0.42), More hurri (0.43), Less hurri (0.15) |
| M 25_34 Day | No (0.32), Yes (0.68) | Sometimes (0.30), Everyday (0.34), Frequently (0.36) | The same (0.23), More hurri (0.69), Less hurri (0.08) |
| M 55_64 Day | No (0.53), Yes (0.47) | Sometimes (0.29), Everyday (0.34), Frequently (0.37) | The same (0.36), More hurri (0.38), Less hurri (0.26) |
| F 15_24 Day | No (0.38), Yes (0.62) | Sometimes (0.33), Everyday (0.36), Frequently (0.31) | The same (0.09), More hurri (0.85), Less hurri (0.06) |
| M 15_24 Day | No (0.52), Yes (0.48) | Sometimes (0.44), Everyday (0.26), Frequently (0.30) | The same (0.17), More hurri (0.79), Less hurri (0.05) |

# SOE - Perception of time

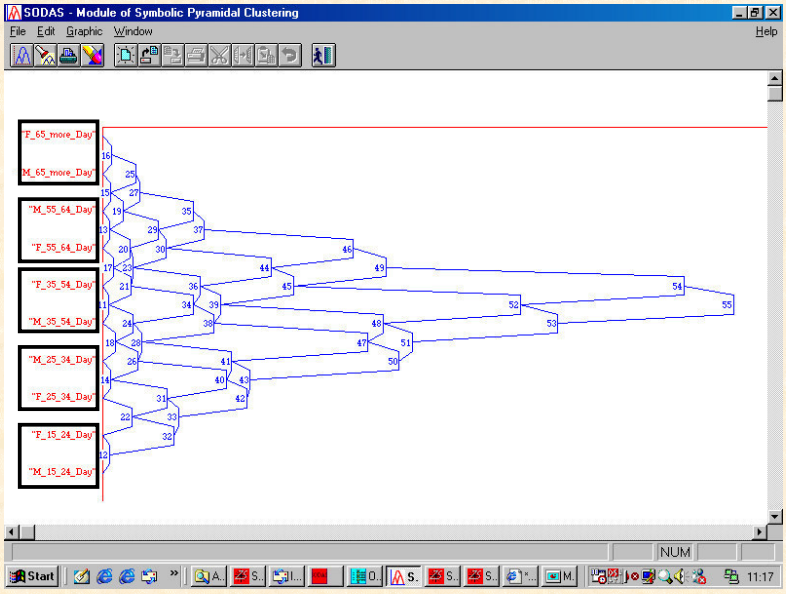



| | Hurried_3years | Hurried_1years |
|---|---|---|
| F 65 more Day | The same (0.57), More hurri (0.19), Less hurri (0.23), Don't know (0.01) | The same (0.67), More hurri (0.17), Less hurri (0.15), Don't know (0.01) |
| M 35_54 Day | The same (0.53), More hurri (0.40), Less hurri (0.07) | The same (0.60), More hurri (0.33), Less hurri (0.06) |
| F 35_54 Day | The same (0.47), More hurri (0.44), Less hurri (0.09) | The same (0.63), More hurri (0.33), Less hurri (0.04) |
| F 25_34 Day | The same (0.39), More hurri (0.55), Less hurri (0.06) | The same (0.55), More hurri (0.40), Less hurri (0.05) |
| M 65 more Day | The same (0.65), More hurri (0.14), Less hurri (0.17), Don't know (0.04) | The same (0.69), More hurri (0.16), Less hurri (0.11), Don't know (0.04) |
| F 55_64 Day | The same (0.56), More hurri (0.31), Less hurri (0.12) | The same (0.68), More hurri (0.26), Less hurri (0.06) |
| M 25_34 Day | The same (0.37), More hurri (0.57), Less hurri (0.06) | The same (0.52), More hurri (0.43), Less hurri (0.05) |
| M 55_64 Day | The same (0.53), More hurri (0.29), Less hurri (0.18) | The same (0.63), More hurri (0.30), Less hurri (0.07) |
| F 15_24 Day | The same (0.17), More hurri (0.72), Less hurri (0.11) | The same (0.42), More hurri (0.45), Less hurri (0.12) |
| M 15_24 Day | The same (0.23), More hurri (0.73), Less hurri (0.04) | The same (0.50), More hurri (0.49), Less hurri (0.01) |

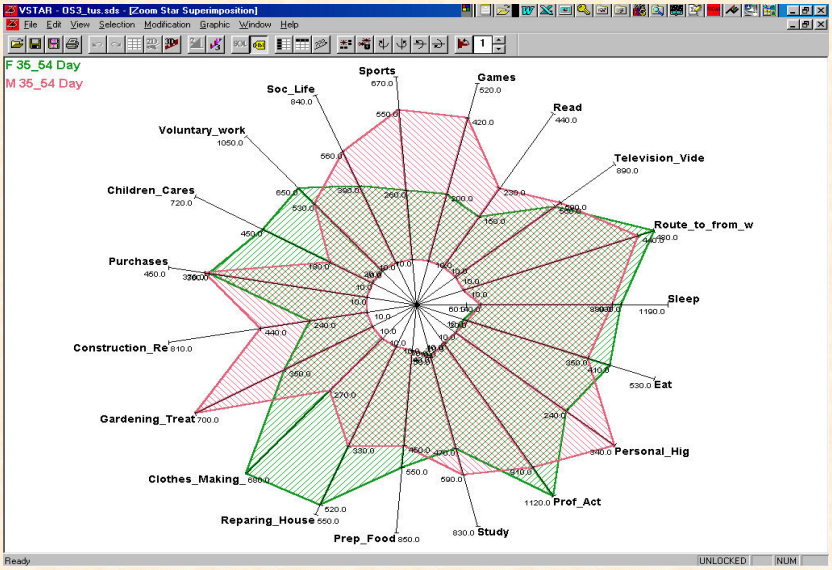

SOE - Editor
SOEditor - OS3B - OS3B.SDS - [Zoom Star - M 35_54 Day]
File Edit View Selection Graphic Window Help
Distribution - M 35_54 Day - Hurried_Often
Sometimes 0.29
Everyday 0.44
Frequently 0.27
M 35_54 Day
Private_Life_Time
Weekend_Time
Work_Time
Available_time
Reduce
Hurried
Hurried_1years
Hurried_3years
Hurried_5years
Hurried_Often
Ready
LOCKED
NUM
Start
11:28


SOE - Editor
SOEditor - OS3B - OS3B.SDS - [Zoom Star - F 35_54 Day]
File Edit View Selection Graphic Window Help
Distribution - F 35_54 Day - Hurried_Often
Sometimes 0.22
Everyday 0.50
Frequently 0.27
F 35_54 Day
Private_Life_Time
Weekend_Time
Work_Time
Available_time
Reduce
Hurried
Hurried_1years
Hurried_3years
Hurried_5years
Hurried_Often
Ready
LOCKED
NUM
Start
11:29

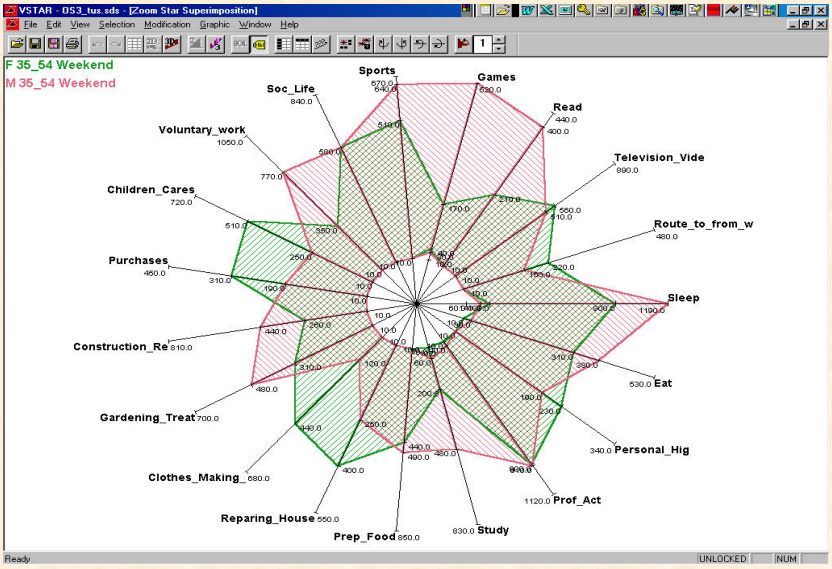
# PYR - Pyramid - hurried



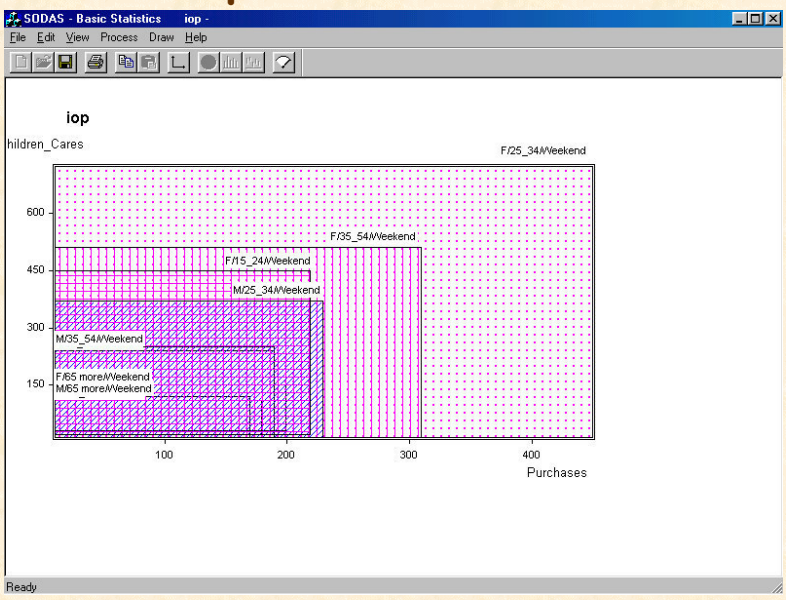

# VIEW - Activities



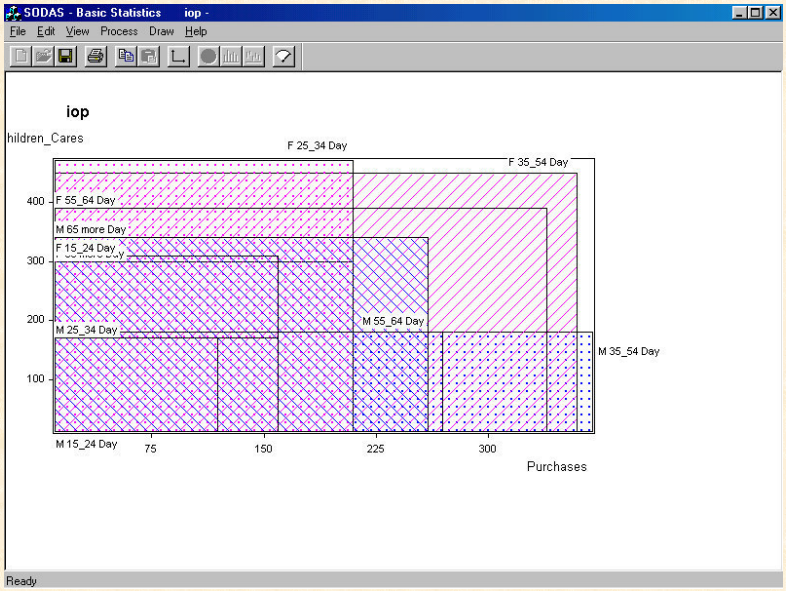
# VIEW - Activities



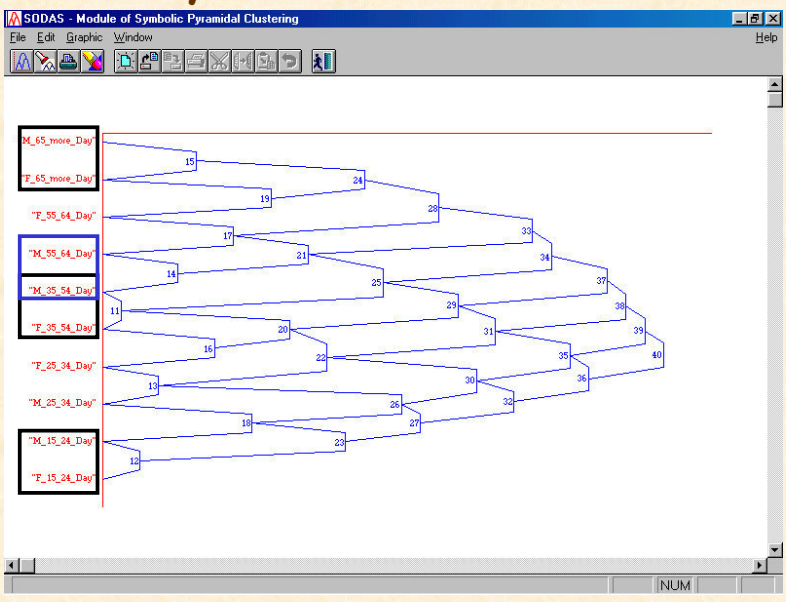

# STAT - Biplot - Activities weekend



# STAT - Biplot - Activities workday




# PYR - Pyramid - activities workday



## Advantages of SDA

- confidentiality of data
- dealing with complex questionnaires
- data analyses using confidence intervals (sampling errors)
- join of independent data files

## The Labour Force Survey

*using confidence intervals*

*Case Study 2*

## DATA

- 2nd quarter 1998 from LFS
- 22660 workers

  *Did you do any work for pay or profit during the reference week (one hour or more) ? (including family workers)*

- processing data:

BLAISE → ASCII → ACCESS → SODAS

## VARIABLES

- **Socio-demographic**
  - age group (ordinal)
  - sex (nominal)
  - marital status (nominal)
  - education level (ordinal)
- **Activity**
  - economic activity (nominal)
  - profession (nominal)
  - searching employment (Boolean)
  - full / part time (Boolean)
  - type of enterprise (nominal)
  - social security (Boolean)
  - over qualifications (Boolean)
  - normal number of hours (Boolean)

## VARIABLES (cont.)

- **Activity (cont.)**
  - previous experience (Boolean)
  - change duration (Boolean)
  - studying or training (Boolean)
  - professional status (nominal)
- **Weight**

**SO**

- sex / age group
- 2*6 = 12

## Analysis with confidence intervals

**$[a_{ij}, b_{ij}]$ - confidence interval**

For each modality (variable) $_j$ and SO $_i$:

$$a_{ij} = p_{ij} - 1.96 * \mathrm{sqrt}(p_{ij} * (1 - p_{ij}) / n_i)$$

$$b_{ij} = p_{ij} + 1.96 * \mathrm{sqrt}(p_{ij} * (1 - p_{ij}) / n_i)$$

$n_i$ - number of the individuals in SO $_i$

$p_{ij}$ - proportion of individuals with the variable $_j$ in SO $_i$

# VARIABLES

-marital status (4)
-education level (4)
-economic activity (12)
-profession (10)
-full / part time (2)
-professional status (5)

TOTAL: 37 var

# VIEW

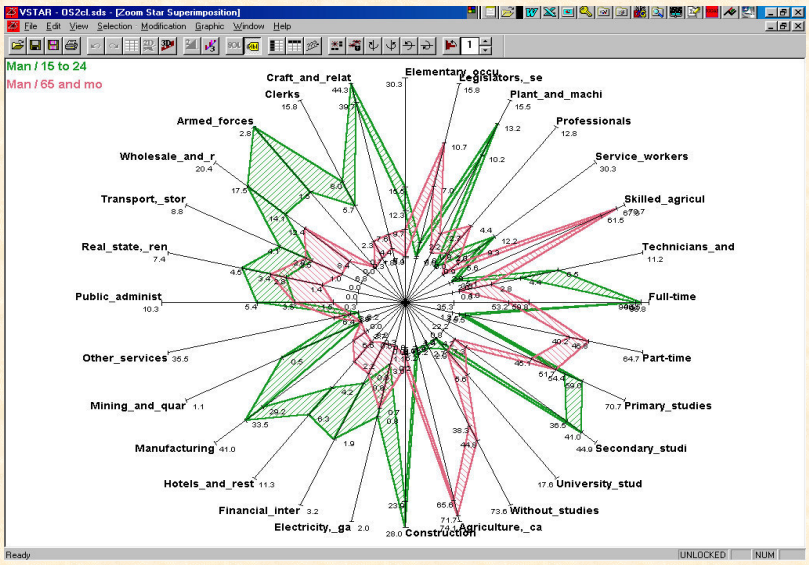
SOEditor - OS2B - OS2.SDS - [Table]

File Edit View Selection Graphic Window Help

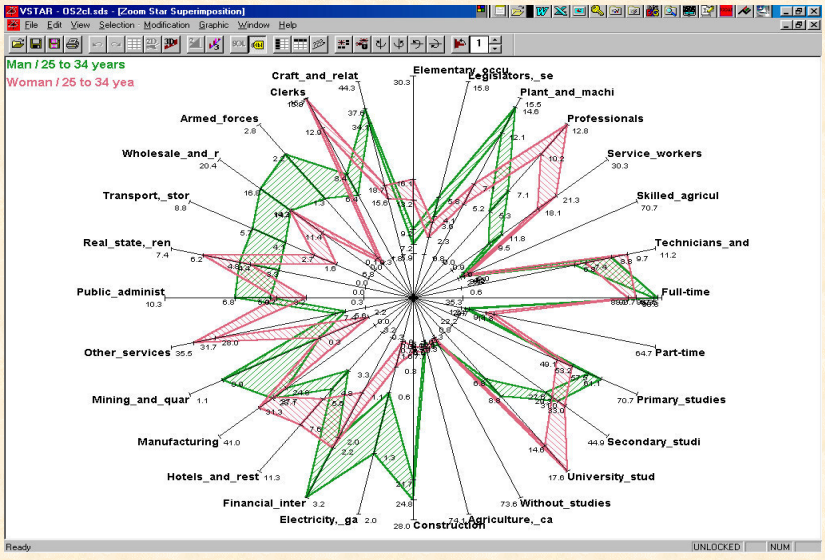
| | Full-time | Part-time | Primary_studies_or_less | Secondary_studies_-_and_profes | University_studies | Without_ |
|---|---|---|---|---|---|---|
| Man / 15 to 24 years old | [ 94.50 : 96.50 ] | [ 3.50 : 5.50 ] | [ 54.40 : 59.00 ] | [ 36.50 : 41.00 ] | [ 1.40 : 2.70 ] | [ 1.80 : |
| Man / 25 to 34 years old | [ 96.30 : 97.60 ] | [ 2.40 : 3.70 ] | [ 57.50 : 61.10 ] | [ 27.60 : 31.00 ] | [ 6.80 : 8.80 ] | [ 2.90 : |
| Man / 35 to 44 years old | [ 97.90 : 98.80 ] | [ 1.20 : 2.10 ] | [ 60.70 : 64.20 ] | [ 23.90 : 27.00 ] | [ 6.70 : 8.60 ] | [ 3.70 : |
| Man / 45 to 54 years old | [ 96.60 : 97.90 ] | [ 2.10 : 3.40 ] | [ 67.10 : 70.70 ] | [ 16.50 : 19.50 ] | [ 6.80 : 8.90 ] | [ 4.30 : |
| Man / 55 to 64 years old | [ 87.20 : 90.10 ] | [ 9.90 : 12.80 ] | [ 63.30 : 67.70 ] | [ 7.10 : 9.70 ] | [ 4.40 : 6.50 ] | [ 18.80 : |
| Man / 65 and more years old | [ 53.20 : 59.80 ] | [ 40.20 : 46.80 ] | [ 45.10 : 51.70 ] | [ 4.20 : 7.30 ] | [ 2.90 : 5.60 ] | [ 38.30 : |
| Woman / 15 to 24 years old | [ 89.10 : 92.30 ] | [ 7.70 : 10.90 ] | [ 46.40 : 51.90 ] | [ 39.50 : 44.90 ] | [ 5.50 : 8.30 ] | [ 1.00 : |
| Woman / 25 to 34 years old | [ 88.20 : 90.70 ] | [ 9.30 : 11.80 ] | [ 49.10 : 53.20 ] | [ 29.30 : 33.00 ] | [ 14.60 : 17.60 ] | [ 1.10 : |
| Woman / 35 to 44 years old | [ 87.40 : 89.90 ] | [ 10.10 : 12.60 ] | [ 56.90 : 60.80 ] | [ 22.10 : 25.40 ] | [ 12.10 : 14.80 ] | [ 3.10 : |
| Woman / 45 to 54 years old | [ 80.00 : 83.40 ] | [ 16.60 : 20.00 ] | [ 61.60 : 65.80 ] | [ 12.10 : 15.10 ] | [ 9.00 : 11.70 ] | [ 10.90 : |
| Woman / 55 to 64 years old | [ 61.60 : 67.00 ] | [ 33.00 : 38.40 ] | [ 43.60 : 49.10 ] | [ 2.90 : 5.10 ] | [ 4.60 : 7.20 ] | [ 41.00 : |
| Woman / 65 and more years old | [ 35.30 : 43.30 ] | [ 56.70 : 64.70 ] | [ 22.20 : 29.30 ] | [ 0.80 : 3.00 ] | [ 1.30 : 3.80 ] | [ 66.10 : |

Ready UNLOCKED NUM

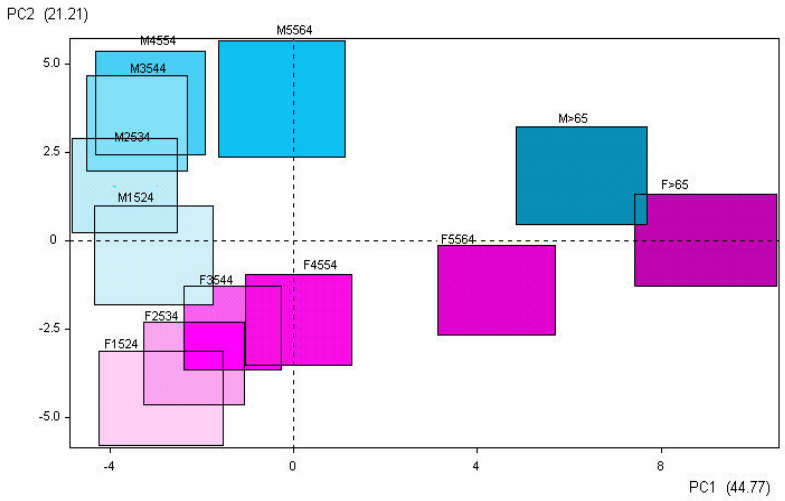
# VIEW


VSTAR - OS2cl.sds - [Zoom Star Superimposition]
File Edit View Selection Modification Graphic Window Help
Man / 15 to 24
Man / 65 and mo
Craft_and_relat
Clerks
Armed_forces
Wholesale_and_r
Transport,_stor
Real_state,_ren
Public_administ
Other_services
Mining_and_quar
Manufacturing
Hotels_and_rest
Financial_inter
Electricity,_ga
Construction
Agriculture,_ca
Without_studies
University_stud
Secondary_studi
Primary_studies
Part-time
Full-time
Technicians_and
Skilled_agricul
Service_workers
Professionals
Plant_and_machi
Legislators,_se
Elementary_occu
Ready
UNLOCKED
NUM

# VIEW


VSTAR - OS2cl.sds - [Zoom Star Superimposition]
File Edit View Selection Modification Graphic Window Help
Man / 25 to 34 years
Woman / 25 to 34 yea
Craft_and_relat
Clerks
Armed_forces
Wholesale_and_r
Transport,_stor
Real_state,_ren
Public_administ
Other_services
Mining_and_quar
Manufacturing
Hotels_and_rest
Financial_inter
Electricity,_ga
Construction
Agriculture,_ca
Without_studies
University_stud
Secondary_studi
Primary_studies
Part-time
Full-time
Technicians_and
Skilled_agricul
Service_workers
Professionals
Plant_and_machi
Legislators,_se
Elementary_occu
Ready
UNLOCKED
NUM

# PCM




# DIV

Xoxim303.lst - WordPad

Fichier Edition Affichage Insertion Format ?

```
PARTITION IN 2 CLUSTERS :
-------------------------:

Cluster 1 (n=9) :

"Man / 15 to 24 years old" "Man / 25 to 34 years old" "Man / 35 to 44 years old"
"Man / 45 to 54 years old" "Man / 55 to 64 years old" "Woman / 15 to 24 years old"
"Woman / 25 to 34 years old" "Woman / 35 to 44 years old" "Woman / 45 to 54 years old"

Cluster 2 (n=3) :

"Man / 65 and more years old" "Woman / 55 to 64 years old" "Woman / 65 and more years old"


PARTITION IN 3 CLUSTERS :
-------------------------:

Cluster 1 (n=2) :

"Man / 15 to 24 years old" "Woman / 15 to 24 years old"

Cluster 2 (n=3) :

"Man / 65 and more years old" "Woman / 55 to 64 years old" "Woman / 65 and more years old"

Cluster 3 (n=7) :

"Man / 25 to 34 years old" "Man / 35 to 44 years old" "Man / 45 to 54 years old"
"Man / 55 to 64 years old" "Woman / 25 to 34 years old" "Woman / 35 to 44 years old"
"Woman / 45 to 54 years old"
```

Pour de l'aide, appuyez sur F1 Num

# DIV

Xoxim303.lst - WordPad

Fichier Edition Affichage Insertion Format ?

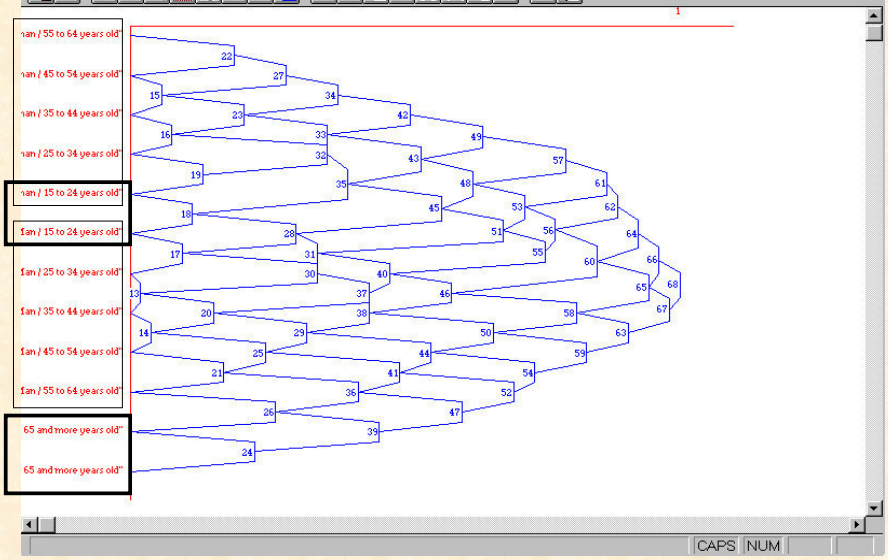
```
PARTITION IN 4 CLUSTERS :

Cluster 1 (n=2) :
"Man / 15 to 24 years old" "Woman / 15 to 24 years old"

Cluster 2 (n=3) :
"Man / 65 and more years old" "Woman / 55 to 64 years old" "Woman / 65 and more years old"

Cluster 3 (n=4) :
"Man / 25 to 34 years old" "Man / 35 to 44 years old" "Man / 45 to 54 years old"
"Man / 55 to 64 years old"

Cluster 4 (n=3) :
"Woman / 25 to 34 years old" "Woman / 35 to 44 years old" "Woman / 45 to 54 years old"


           +----  Classe 1 (Ng=2)
           !
      !----2- [Married <= 46.200001]
      !    !
      !    !    +----  Classe 3 (Ng=4)
      !    !    !
      !    !----3- [Other_services <= 19.075000]
      !         !
      !         +----  Classe 4 (Nd=3)
      !
 !----1- [Part-time <= 27.000000]
      !
      +----  Classe 2 (Nd=3)
```

Pour de l'aide, appuyez sur F1 Num

# PYR

CEREMADE - Module of Symbolic Pyramidal Clustering

File Edit Classification Options Window Help




CAPS NUM

# FDA

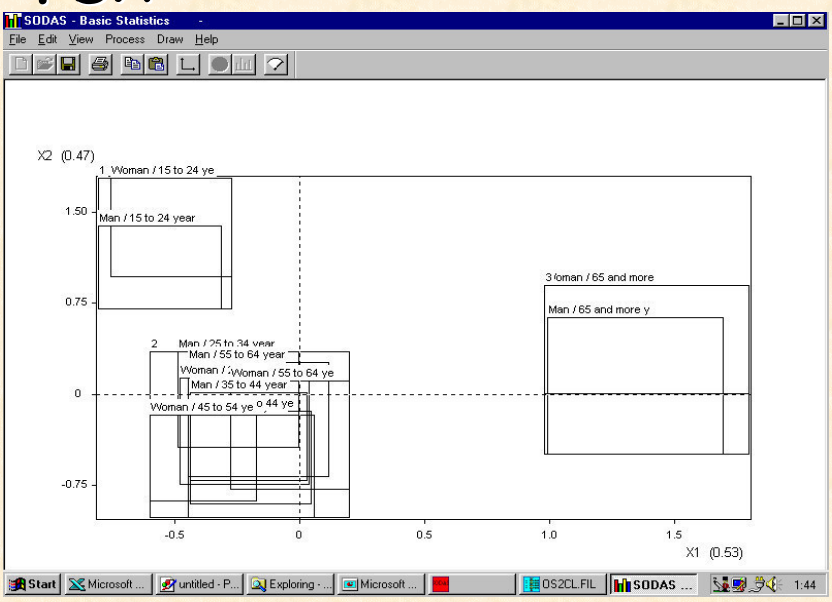
Xoys2e02.lst - WordPad

Fichier Edition Affichage Insertion Format ?

```
Classification table:
-----------------------------------------------
Test\Classes          |1        |2        |3
-----------------------------------------------
"Man / 15 to 24 year| 1         0         0
"Man / 25 to 34 year| 0         1         0
"Man / 35 to 44 year| 0         1         0
"Man / 45 to 54 year| 0         1         0
"Man / 55 to 64 year| 0         1         0
"Man / 65 and more y| 0         0         1
"Woman / 15 to 24 ye| 1         0         0
"Woman / 25 to 34 ye| 0         1         0
"Woman / 35 to 44 ye| 0         1         0
"Woman / 45 to 54 ye| 0         1         0
"Woman / 55 to 64 ye| 0         1         0
"Woman / 65 and more| 0         0         1
-----------------------------------------------


Classification Summary :
-----------------------------------------------
Class\Class           |1        |2        |3
-----------------------------------------------
1                     | 2         0         0
2                     | 0         8         0
3                     | 0         0         2
-----------------------------------------------
Classification Ratio : 100.0%
-----------------------------------------------
```

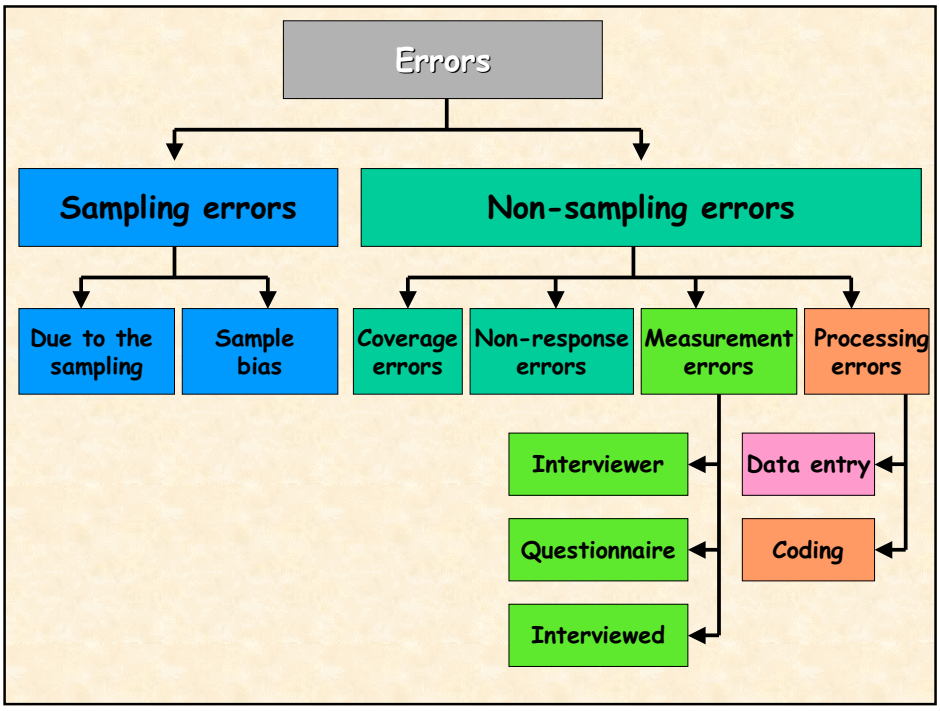
Pour de l'aide, appuyez sur F1 Num

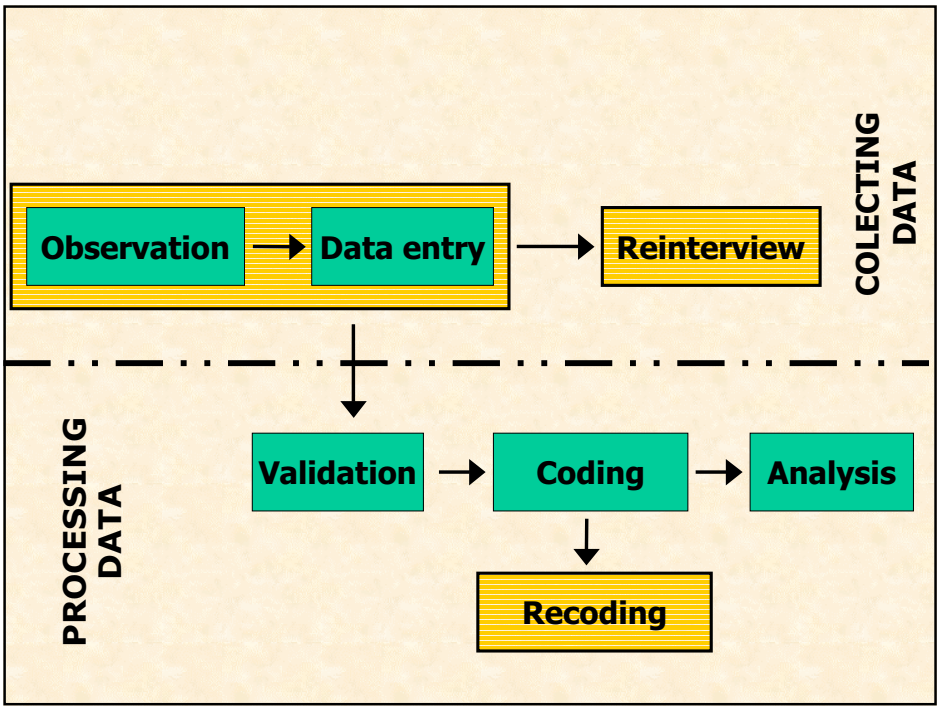
# FDA



# The Quality Control in Statistical Production

*Case Study 3*

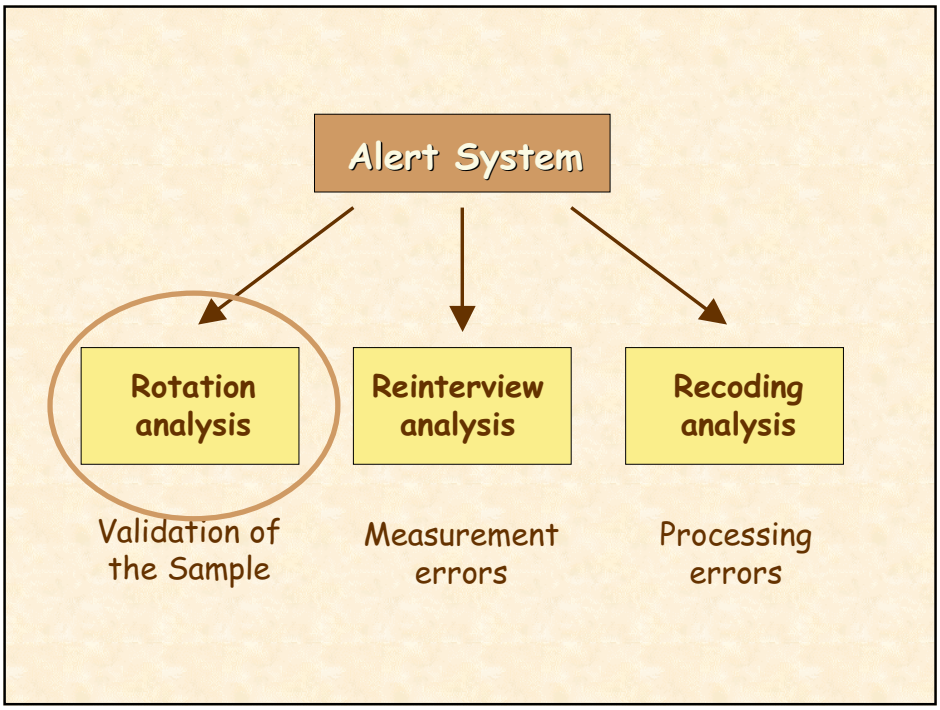





## Quality Control

- in a **quantitative** and **univariate** form at micro-data level, using some indicators already presented in several bibliography (consistency index)

- in a **qualitative** form, using **multivariate** techniques on the aggregation of the registers, according to convenient *criteria* for each of the different analysis to be done. It is a control system based on symbolic data analysis using SODAS software




## Quality control for non-sampling errors

- validation of the sample in panels with rotation
- measurement errors - reinterview analysis
- processing errors - recoding analysis

## DATA

- **Observed Data**

  All employed persons from the Portuguese LFS 1st and 2nd quarters of 2000 - region of "Lisboa e Vale do Tejo" (8669 records)

- **Processing data**

  BLAISE → ASCII → ACCESS → SODAS

- **Symbolic Objects**

  on rotation of the sample (6*2=12 SO)

## DATA

**Sampling frame**

Master Sample - list of housing units

(dwellings)

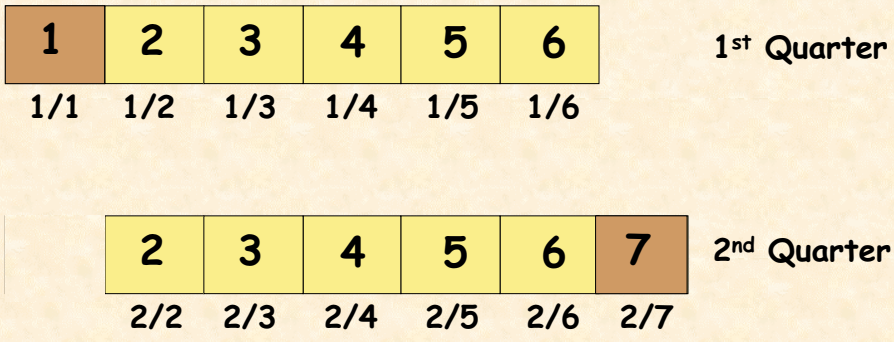
**sampling unit** - dwelling

**observation unit** - individual

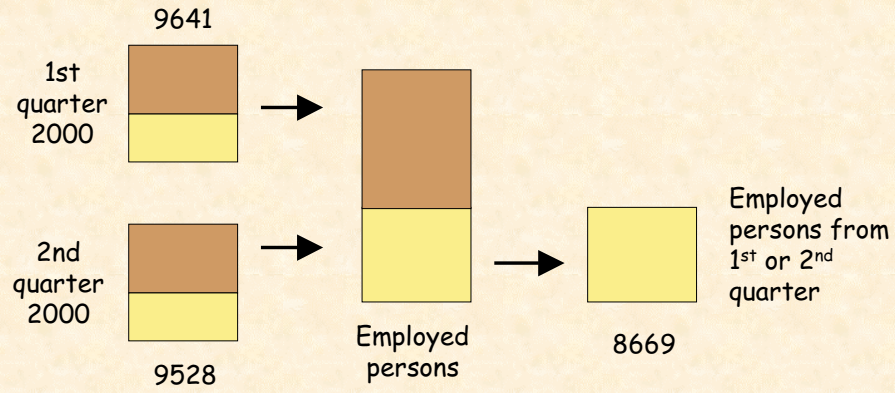
# VARIABLES

- -socio-demographic:
  - sex
  - age group
  - marital status
  - education level
- -activity variables:
  - occupation
  - economic activity
  - professional status
  - part/full time
  - number of hours worked group
- -sampling weight

## Rotation Analysis



## Rotation Analysis




## Access Table

| | key | rot | sex | age_group | marital_status | economic_activity |
|---|---|---|---|---|---|---|
| ▶ | 1510203011 | 1 / 2 | Male | 55 <= age <= 64 | Married | Wholesale and retail trade, repairs |
| | 1510204112 | 1 / 6 | Male | 35 <= age <= 44 | Single | Construction |
| | 1511203211 | 1 / 1 | Male | 35 <= age <= 44 | Married | Manufacturing |
| | 1511203511 | 1 / 2 | Male | 35 <= age <= 44 | Married | Manufacturing |
| | 1511203512 | 1 / 2 | Female | 25 <= age <= 34 | Married | Other Services |
| | 1511203711 | 1 / 2 | Male | 45 <= age <= 54 | Married | Manufacturing |
| | 1511203712 | 1 / 2 | Female | 45 <= age <= 54 | Married | Manufacturing |
| | 1511203713 | 1 / 2 | Male | 15 <= age <= 24 | Single | Manufacturing |
| | 1511203714 | 1 / 2 | Female | 15 <= age <= 24 | Single | Manufacturing |
| | 1511203811 | 1 / 2 | Female | 15 <= age <= 24 | Single | Manufacturing |
| | 1511203911 | 1 / 6 | Male | 35 <= age <= 44 | Married | Manufacturing |
| | 1511203912 | 1 / 6 | Female | 35 <= age <= 44 | Married | Manufacturing |
| | 1511204111 | 1 / 3 | Female | 35 <= age <= 44 | Divorced or Separated | Other Services |
| | 1511204411 | 1 / 4 | Male | 45 <= age <= 54 | Married | Manufacturing |
| | 1511204412 | 1 / 4 | Female | 45 <= age <= 54 | Married | Manufacturing |
| | 1511204413 | 1 / 4 | Male | 15 <= age <= 24 | Single | Wholesale and retail trade, repairs |
| | 1511204511 | 1 / 4 | Male | 55 <= age <= 64 | Married | Manufacturing |
| | 1511204611 | 1 / 5 | Female | 45 <= age <= 54 | Divorced or Separated | Other Services |
| | 1511204711 | 1 / 5 | Female | 35 <= age <= 44 | Married | Other Services |

key SO Var.

Symbolic table

SOEditor - AAA.FIL - ROTTOT.SDS - [Table]

File Edit View Selection Graphic Window Help

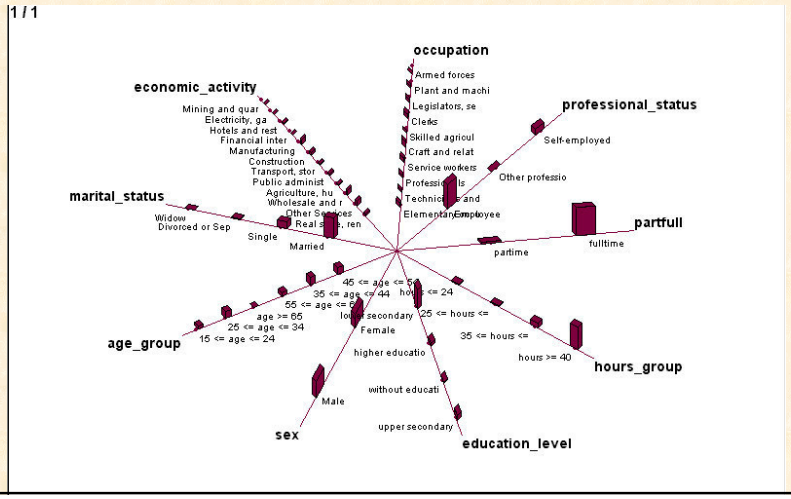
| | sex | age_group |
|---|---|---|
| 2 / 6 | Female (0.45), Male (0.55) | 45 <= age <= 54 (0.22), 35 <= age <= 44 (0.28), 55 <= age <= 64 (0.08), age >= 65 (0.02), 25 <= age <= 3 |
| 2 / 5 | Female (0.46), Male (0.54) | 45 <= age <= 54 (0.22), 35 <= age <= 44 (0.25), 55 <= age <= 64 (0.14), age >= 65 (0.03), 25 <= age <= 3 |
| 2 / 3 | Female (0.47), Male (0.53) | 45 <= age <= 54 (0.25), 35 <= age <= 44 (0.22), 55 <= age <= 64 (0.12), age >= 65 (0.02), 25 <= age <= 3 |
| 2 / 4 | Female (0.45), Male (0.55) | 45 <= age <= 54 (0.22), 35 <= age <= 44 (0.22), 55 <= age <= 64 (0.14), age >= 65 (0.04), 25 <= age <= 3 |
| 2 / 7 | Female (0.47), Male (0.53) | 45 <= age <= 54 (0.23), 35 <= age <= 44 (0.24), 55 <= age <= 64 (0.12), age >= 65 (0.04), 25 <= age <= 3 |
| 1 / 1 | Female (0.45), Male (0.55) | 45 <= age <= 54 (0.22), 35 <= age <= 44 (0.26), 55 <= age <= 64 (0.13), age >= 65 (0.03), 25 <= age <= 3 |
| 2 / 2 | Female (0.47), Male (0.53) | 45 <= age <= 54 (0.23), 35 <= age <= 44 (0.24), 55 <= age <= 64 (0.13), age >= 65 (0.03), 25 <= age <= 3 |
| 1 / 5 | Female (0.46), Male (0.54) | 45 <= age <= 54 (0.21), 35 <= age <= 44 (0.25), 55 <= age <= 64 (0.14), age >= 65 (0.02), 25 <= age <= 3 |
| 1 / 2 | Female (0.47), Male (0.53) | 45 <= age <= 54 (0.23), 35 <= age <= 44 (0.24), 55 <= age <= 64 (0.13), age >= 65 (0.03), 25 <= age <= 3 |
| 1 / 3 | Female (0.47), Male (0.53) | 45 <= age <= 54 (0.25), 35 <= age <= 44 (0.21), 55 <= age <= 64 (0.12), age >= 65 (0.02), 25 <= age <= 3 |
| 1 / 6 | Female (0.44), Male (0.56) | 45 <= age <= 54 (0.21), 35 <= age <= 44 (0.28), 55 <= age <= 64 (0.08), age >= 65 (0.02), 25 <= age <= 3 |
| 1 / 4 | Female (0.45), Male (0.55) | 45 <= age <= 54 (0.23), 35 <= age <= 44 (0.21), 55 <= age <= 64 (0.13), age >= 65 (0.03), 25 <= age <= 3 |

```
2 / 7 =
        sex = Female (0.47), Male (0.53)
And     age_group = 45 <= age <= 54 (0.23), 35 <= age <= 44 (0.24), 55 <= age <= 64 (0.12), age >= 65 (0.04) 25 <= age
And     marital_status = Married (0.69), Single (0.22), Divorced or Sep (0.05), Widow (0.03)
And     economic_activity = Real state, ren (0.07), Other Services (0.21), Wholesale and r (0.17), Agriculture hu (0.06),
And     occupation = Elementary occu (0.16), Technicians and (0.10), Professionals (0.08), Service workers (0.15), Craft
And     professional_status = Employee (0.79), Other professio (0.02), Self-employed (0.19)
And     partfull = partime (0.11), fulltime (0.89)
And     hours_group = hours <= 24 (0.08), 25 <= hours <= (0.04), 35 <= hours <= (0.18), hours >= 40 (0.70)
And     education_level = lower secondary (0.66), higher educatio (0.13), without educati (0.06), upper secondary (0.15)
And     rot_cl = 1
```

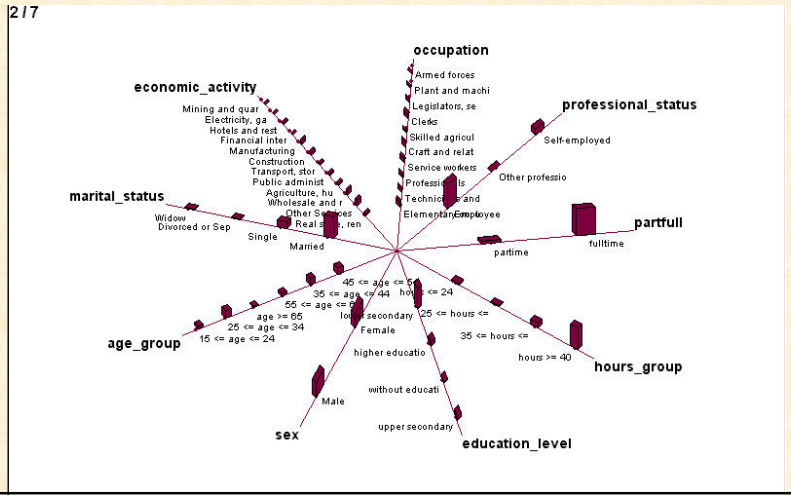
SOL

Symbolic Object Language

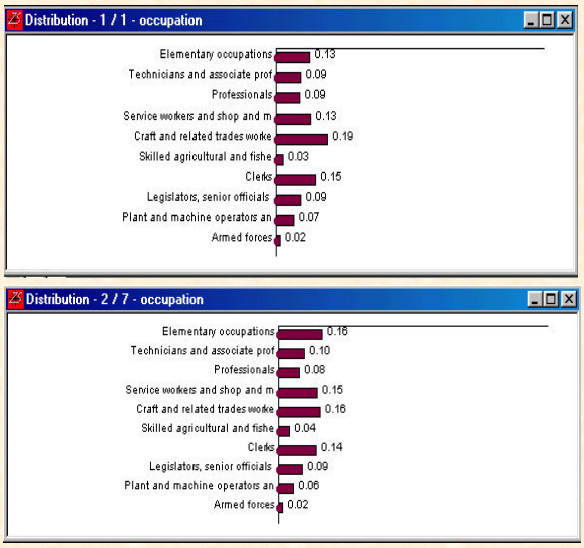
# Editor - Rotation 1/1



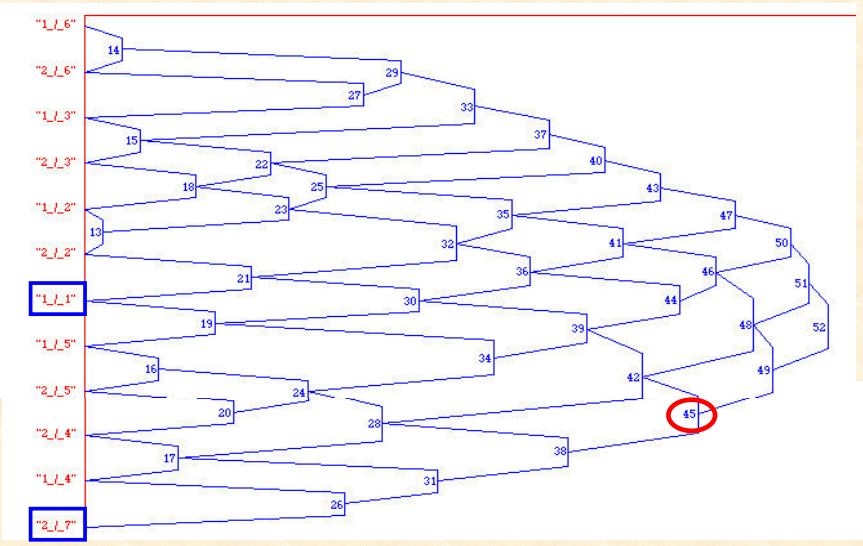
# Editor - Rotation 2/7




# Editor - Occupation 1/1 and 2/7



# Pyramid




# Factorial Discriminant Analysis

SOEditor - AAA.FIL - ROTTOT.SDS - [Table]

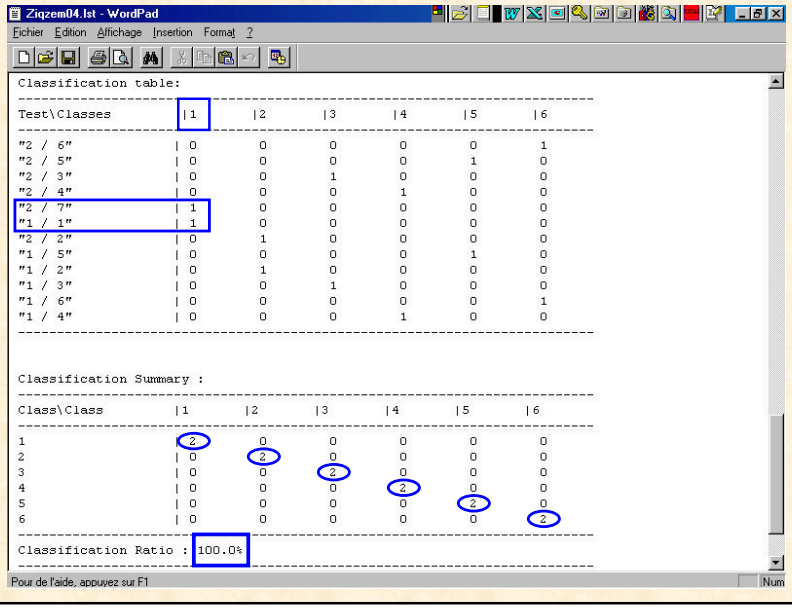
File Edit View Selection Graphic Window Help

| | education_level | rot_cl |
|---|---|---|
| 2 / 6 | lower secondary (0.66), higher educatio (0.14), without educati (0.05), upper secondary (0.15) | 6 |
| 2 / 5 | lower secondary (0.66), higher educatio (0.12), without educati (0.05), upper secondary (0.17) | 5 |
| 2 / 3 | lower secondary (0.63), higher educatio (0.15), without educati (0.06), upper secondary (0.16) | 3 |
| 2 / 4 | lower secondary (0.67), higher educatio (0.11), without educati (0.06), upper secondary (0.15) | 4 |
| 2 / 7 | lower secondary (0.66), higher educatio (0.13), without educati (0.06), upper secondary (0.15) | 1 |
| 1 / 1 | lower secondary (0.64), higher educatio (0.12), without educati (0.06), upper secondary (0.18) | 1 |
| 2 / 2 | lower secondary (0.62), higher educatio (0.15), without educati (0.06), upper secondary (0.17) | 2 |
| 1 / 5 | lower secondary (0.66), higher educatio (0.12), without educati (0.06), upper secondary (0.16) | 5 |
| 1 / 2 | lower secondary (0.61), higher educatio (0.15), without educati (0.06), upper secondary (0.18) | 2 |
| 1 / 3 | lower secondary (0.62), higher educatio (0.15), without educati (0.06), upper secondary (0.17) | 3 |
| 1 / 6 | lower secondary (0.66), higher educatio (0.14), without educati (0.05), upper secondary (0.15) | 6 |
| 1 / 4 | lower secondary (0.68), higher educatio (0.11), without educati (0.06), upper secondary (0.15) | 4 |

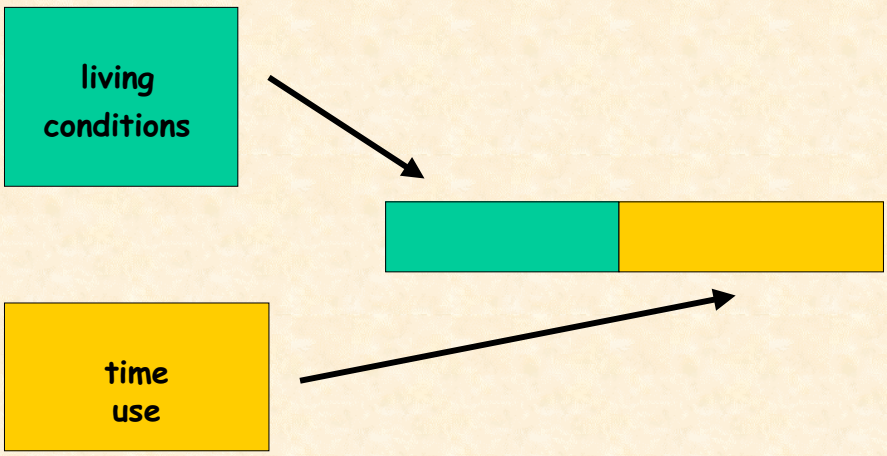
SO

Class var

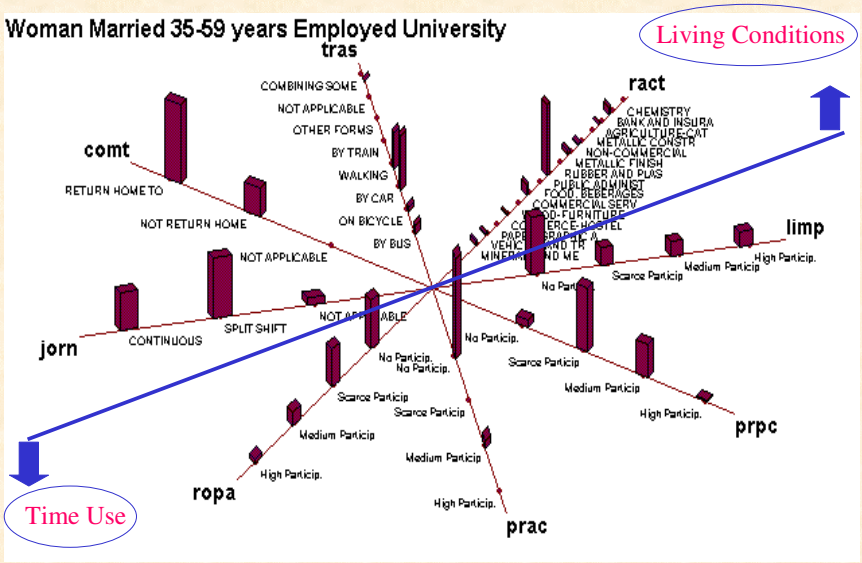
# Factorial Discriminant Analysis




# Joining of independent surveys



# Joining of independent surveys

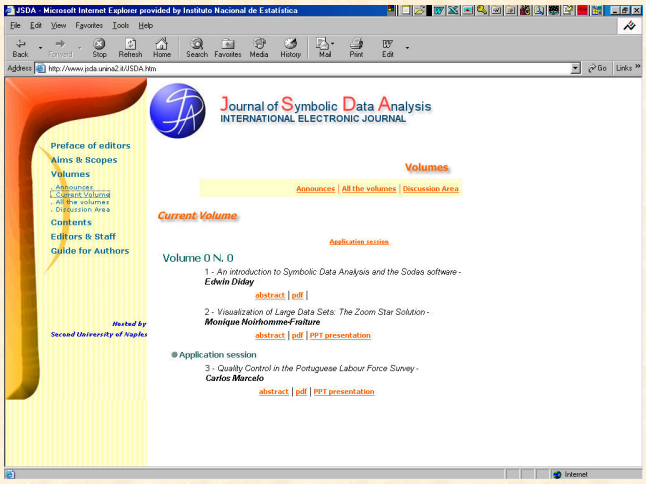



# Dissemination

**One source of information related with Symbolic Data Analysis is the Journal of Symbolic Data Analysis, an international electronic journal where research articles and data applications can be found**

Journal of Symbolic Data Analysis
INTERNATIONAL ELECTRONIC JOURNAL

**http://www.jsda.unina2.it**



SODAS software is a useful tool for:

- Manipulating
- Joining
- Visualizing
- Comparing
- Analyzing
- Explaining

COMPLEX

official statistical data EASIER !!!

## References


- Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data - BOCK, H.H.; DIDAY, E.; Springer, 2000
- SODAS Reference Manual
- **Download SODAS I and several information:**
  www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm
- **ASSO site**
  www.assoproject.be
- **Journal of Symbolic Data Analysis**
  www.jsda.unina2.it