KNOWLEDGE DISCOVERY FROM SYMBOLIC DATA AND
THE SODAS SOFTWARE

Edwin DIDAY
University Paris 9 Dauphine, Ceremade. Pl. Du Mle de L. de Tassigny. 75016 Paris,
FRANCE

Abstract. The data descriptions of the units are called "symbolic" when they are more complex than the standard ones due to the fact that they contain internal variation and are structured. Symbolic data happen from many sources, for instance in order to aggregate huge Relational Data Bases by their underlying concepts. "Extracting knowledge" means getting explanatory results, that why, "symbolic objects" are introduced and studied in this paper. They model concepts and constitute an explanatory output for data analysis. Moreover they can be used in order to define queries of a Relational Data Base and propagate concepts between Data Bases. We define "Symbolic Data Analysis" (SDA) as the extension of standard Data Analysis to symbolic data tables as input in order to find symbolic objects as output. Any SDA is based on four spaces: the space of individuals, the space of concepts, the space of descriptions modelling individuals or classes of individuals, the space of symbolic objects modelling concepts. Based on these four spaces, new problems appear as the quality, robustness and reliability of the approximation of a concept by a symbolic object, the symbolic description of a class, the consensus between symbolic descriptions etc.. In this paper we give an overview on recent development on SDA. We present some tools and methods of SDA and introduce the SODAS software prototype (issued from the work of 17 teams of nine countries involved in an European project of EUROSTAT).

Introduction
As input, when large data sets are aggregated into smaller more manageable data sizes we need more complex data tables called "symbolic data tables" because a cell of such data table does not necessarily contain as usual, a single quantitative or categorical value.
In a symbolic data table, a cell can contain, a distribution (Schweitzer (1984) says that "distributions are the number of the future"!), or intervals, or several values linked by a taxonomy and logical rules, etc.. The need to extend standard data analysis methods (exploratory, clustering, factorial analysis, discrimination,...) to symbolic data table is increasing in order to get more accurate information and summarise extensive data sets contained in Data Bases.
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.

Our first aim is to extend standard data analysis to second order individuals. For instance, in a census of a country, each individual of each region is described by a set of numerical or categorical variables given in several relations of a Data Base. Such an individual is considered as a "first order individual". In order to study the regions considered as "second order individuals", we can describe each of them in summarising the values taken by its inhabitants, by inter-quartile intervals, or subsets of categorical values, or histograms or probability distributions, etc. depending on the concerned 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. 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 a "concept" or a "physical entity" of the real world. A "symbolic object" is defined by its "intent" which contains a way of finding its "extent". For instance, the description of the inhabitant 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 4, for a formal definition): a(w) = [age(w) ∈ [30, 35]] ∧ [Number of children(w) ≤ 2], gives the intent of a class of individuals 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 "w" satisfies this intent (i.e. belongs to the "extent") if his age is between 30 and 35 years old and he has 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 Data Base and therefore, to propagate the
"concepts" that they describe from one data base to another database (for instance, from a
country to another country).

Symbolic objects model "concepts" but what do we call a "concept"?

A "concept" is generally defined by a set of properties called "intent" or "comprehension" and
a set of individuals which satisfy these properties called "extent" or "comprehension". By
using "universal ideas" for "concept", 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" 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 recognising 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
modelled 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 problem will be discussed in § 4.

**Concepts: four tendencies.**

In the Aristotelian tradition, concepts are characterised by logical conjunction of properties. In
the Adansonian tradition (Adanson (1727-1806) was a French naturalist very much ahead of
his time), a concept is characterised 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.

- 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). 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.

- The 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 (1991,

**Symbolic Data Analysis is born from which influence?**
There was a simultaneous influence of several fields:
- standard exploratory data analysis (Tukey (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 modelled by a symbolic object.
- Classification in Data Analysis where a class can be modelled by a symbolic object.

In all these fields a natural question arose: how does one obtain classes and their description?

Historically, we may say briefly that there are 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. 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).
The second tendency, 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) an ascending pyramid produces monotheistic classes.

The third tendency consists in looking directly for classes and their 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 representation in such a way that they "fit" together as well as possible. This approach has been used with several kinds of inter-class structure (partitions, hierarchies, ...) and 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".

**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. Schecktman (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) and the section 9.3 in this book. In the case where the individuals are described by symbolic data, Conruyt (1993) in the case of structured data, Ciaimi, E. Diday, J. Lebbe, E. Périnel, R. Vigne (1995), Périnel (1996), 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)) and Gettler-Summa in this book. 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) 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 provides also a symbolic object associated to each obtained class 
(see chapter 11 in Bock , Diday (2000)). Stéphan (1998) for extracting symbolic objects from 
a data base(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 E. (2000)). B. Tang (1998) for extending 
Factorial Correspondence Analysis and O. Rodriguez (2000) for extending regression and 
Multidimensional Scaling to interval data.

2) The input of a symbolic data analysis: a "symbolic data table"

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

<table>
<thead>
<tr>
<th>WAGES</th>
<th>TOWN</th>
<th>SOCIO-ECONOMIC GROUP</th>
</tr>
</thead>
<tbody>
<tr>
<td>{3.5}</td>
<td>{London}</td>
<td>{Personal of service}</td>
</tr>
<tr>
<td>[3, 8]</td>
<td>{Paris, London}</td>
<td></td>
</tr>
<tr>
<td>{3.1, 4.6, 7.2}</td>
<td></td>
<td>{0.1 Manager, 0.6 Manual, ...}</td>
</tr>
<tr>
<td>[(0.4) [2,3[, (0.6) [3, 8]]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. A "symbolic data table": each cell contains an example of "symbolic data".

3) Sources of Symbolic Data:

Symbolic data are generated when we summarise huge sets of data. The need of such summary can appear in different ways, for instance, from any query to a data base which induces categories and descriptive variables. These categories can be for instance, simply the towns 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 Data Bases, 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 Data Base 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.

<table>
<thead>
<tr>
<th>Delivery</th>
<th>Supplier</th>
<th>Company</th>
<th>Town</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liv1</td>
<td>F1</td>
<td>CNET</td>
<td>Paris</td>
</tr>
<tr>
<td>Liv2</td>
<td>F2</td>
<td>MATRA</td>
<td>Toulouse</td>
</tr>
<tr>
<td>Liv3</td>
<td>F3</td>
<td>EDF</td>
<td>Clamart</td>
</tr>
<tr>
<td>Liv4</td>
<td>F1</td>
<td>CNET</td>
<td>Lannion</td>
</tr>
<tr>
<td>Liv5</td>
<td>F3</td>
<td>EDF</td>
<td>Clamart</td>
</tr>
</tbody>
</table>

Table 1 Relation "Delivery"

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

<table>
<thead>
<tr>
<th>Supplying</th>
<th>Supplier</th>
<th>Town</th>
</tr>
</thead>
<tbody>
<tr>
<td>FT1</td>
<td>F1</td>
<td>Paris</td>
</tr>
<tr>
<td>FT2</td>
<td>F2</td>
<td>Toulouse</td>
</tr>
<tr>
<td>FT3</td>
<td>F1</td>
<td>Lannion</td>
</tr>
<tr>
<td>FT4</td>
<td>F3</td>
<td>Clamart</td>
</tr>
<tr>
<td>FT5</td>
<td>F3</td>
<td>Clamart</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th>Supplier</th>
<th>Company</th>
<th>Supplying</th>
<th>Town</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>CNET</td>
<td>FT1, FT3</td>
<td>_ Paris, _ Lannion</td>
</tr>
<tr>
<td>F2</td>
<td>MATRA</td>
<td>FT2</td>
<td>Toulouse</td>
</tr>
<tr>
<td>F3</td>
<td>EDF</td>
<td>FT4, FT5</td>
<td>Clamart</td>
</tr>
</tbody>
</table>

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 _ means that the town of the supplier F1 is Paris or Lannion with a frequency equal to _.

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".

4) Main output of Symbolic Data Analysis algorithms:

4) Symbolic objects: definition and properties.

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 = "∈" and "a" is the mapping: Ω → {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 Ω be a set of individuals, D a set containing descriptions of individuals or of a class of individuals, « y » a mapping defined from Ω into D which associates to each w ∈ Ω a description d ∈ 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 \( L = \{\text{true, false}\} \), in this case \([d' R d] = \text{true}\) 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, \( R \in \{=, \equiv, \leq, \subseteq\} \) 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 \( \Omega \) 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 \([y(w) R d] \in L = \{\text{true, false}\}\). In this case, if \( y(w) = (y_1, ..., y_p) \), the \( y_i \) are of type (a) to (d), defined in section 1.

  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 \( \vee \) has the standard logical meaning and \( R_i = \subseteq \). If \( y(w) = (\text{colour}(w), \text{height}(w)) \), \( d = (\{\text{red, blue, yellow}\}, [10,15]) \) = \([d_1, d_2]\), \( \text{colour}(u) = \{\text{red, yellow}\} \), \( \text{height}(u) = \{21\} \), then

  \( a(u) = [\text{colour}(u) \subseteq \{\text{red, blue, yellow}\}] \vee [\text{height}(u) \subseteq [10,15]] = \text{true} \wedge \text{false} = \text{true} \).

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

  Example:

  Let be \( a(u) = [y(u) R d] \) where for instance \( R: [d' R d] = \text{Max}_{i=1,2} [d'_i, R_i, d_i] \). 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 \( d' = r \) and \( d = q \) of \( k \) values by:

\[
r R_d = \sum_{j=1}^{k} r_j q_j e^{(r_j - \min(r_j, q_j))}.
\]

By analogy with the Boolean case we denote \( [d' R d] = \lor^* \) where \( \lor^* = \max \). 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 \( d = \{(0.2)_{12}, (0.8)_{[20 \ldots 28]}\}, \{(0.4)_{employee}, (0.6)_{worker}\} \) by:

\[
a(u) = [\text{age}(u) \lor_R (0.2)_{12}, (0.8)_{[20 \ldots 28]}] \lor^* [\text{SPC}(u) \lor_R (0.4)_{employee}, (0.6)_{worker}] 
\]

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 \( R_i \) depends on this choice, (see Diday (1995), for instance).

**Syntax of symbolic objects in the case of "assertions":**

If the initial data table contains \( p \) variables we denote \( y(w) = (y_1(w), \ldots, y_p(w)) \), \( D = (D_1, \ldots, D_p) \), \( d \in D : d = (d_1, \ldots, d_p) \) and \( R' = (R_1, \ldots, R_p) \) 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

\[
[d' R d] = \land_{i=1}^{p} [d'_i R d_i] \quad \text{where } \land \text{ has the standard logical meaning and } "a" \text{ is defined by: }
\]

\[
a(w) = [y(w) R d] \text{ in the Boolean case. Notice that considering the expression } 
\]

\[
a(w) = \land_{i=1}^{p} [y_i(w) R_i d_i] \text{ we are able to define the symbolic object } s = (a, R, d). \text{ Hence, we can say that this explanatory expression defines a symbolic object called "assertion".}
\]

For example, a Boolean assertion is:

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

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

\[
a(u) = [y(u) R d] \text{ with } [d' R d] = f([\{y_i(w) R_i d_i\}]_{i=1,p}) \text{ where for instance,}
\]

\[
f([\{y_i(w) R_i d_i\}]_{i=1,p}) = \prod_{i=1}^{2} [d'_i R_i d_i] \quad \text{where in case of probability distributions, the "matching" is defined for two discrete density distributions } d'_i = \tau = (r_1, \ldots, r_k) \text{ and } d_i = q = (q_1, \ldots, q_k) \text{ of } k \text{ values by: } r R_i q = \sum_{j=1}^{k} r_j q_j e^{(r_j - \min(r_j, q_j))}.
\]

By analogy with the Boolean case we denote \( [d' R d] = \land^*_{i=1, 2} p_i [d'_i R_i d_i] \) where the meaning of "\( \land^* \)" 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:

\[
a(w) = [\text{age}(w) \cdot R_1 \{(0.2)12, (0.8) [20,28]\}] \land [\text{SPC}(w) \cdot R_2 \{(0.4)\text{employee}, (0.6)\text{worker}\}]
\]

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

**Other possible classes of symbolic objects:** If for instance the mapping "a" is of the following kind: \( a(w) = [h_e(y(w)) \cdot h_j(R) \cdot h_i(d)] \), different classes of symbolic objects may be defined depending on the choice of \( h_e, h_j \) 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 a filter on the extension:**

We associate to each town a symbolic object defined by \( a(w) = [h_e(y(w)) \cdot R \cdot d] \) 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 \( [h_e(y(w)) \cdot R \cdot d] = 0 \) for any description \( d \).

**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: \( s_1 \prec_r s_2 \) iff \( d_1 \succ_r d_2 \).

If \( R \) is such that \( [d \cdot R \cdot d'] = \text{true} \) implies \( d \cdot R \cdot d' \), then \( \text{Ext}(s1) \subseteq \text{Ext}(s2) \) if \( s1 \prec_r s2 \). If \( R \) is such that \( [d \cdot R \cdot d'] = \text{true} \) implies \( d' \cdot R \cdot d \) then \( \text{Ext}(s2) \subseteq \text{Ext}(s1) \) if \( s1 \prec_r s2 \).

**Tools for symbolic objects:** Tools between symbolic objects (Diday (1995)) are needed such as similarities (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).

**Description of a class of individuals:**

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 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)). For instance, in a top-down clustering tree where at each step a splitting variable is chosen 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 distances 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 splitting 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 initial set of individuals in order to summarize 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.

**Underlying structures of symbolic objects: a generalised conceptual lattice**

Under some assumptions on the choice of R and T (for instance $T \equiv \text{Max}$ if $R \equiv \leq$ and $T \equiv \text{Min}$ if $R \equiv \geq$) it can be shown that the underlying structure of a set of symbolic objects is a Galois lattice (Diday (1991), Brito (1994), Diday, Emilion (1995), (1997), Polaillon, Diday (1997), Polaillon (1998), Bock, Diday (2000)), where the vertices are closed sets defined thereunder 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 $T_{c \in C} y(c) = \cup_{c \in C} y(c)$, $R \equiv \subseteq$, $y(u) = \{\text{pink, blue}\}$, $C = \{c, c'\}$, $y(c) = \{\text{pink, red}\}$, $y(c') = \{\text{blue, red}\}$, then $a(u) = [y(w) \ R \ T_{c \in C} y(c)] = \{\text{pink, blue}\} \subseteq \{\text{pink, red}\} \cup \{\text{blue, red}\} \} = \{\text{pink, red, blue}\} = \text{true and } u \in \text{ Ext } (s)$. 

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.

Finally we can summarize the mathematical framework of a symbolic data analysis in the following way (figure 1):

In order to see how much a given symbolic object is characteristic of a class $A$, an hypergeometric distribution can be used. Let $N$ be the size of $\Omega'$, $n$ the size of $A$, $p = \text{Ext}(s/\Omega') / N$ the proportion in $\Omega'$ of individuals belonging in the extent of $s$, $X$ a random variable whose value at each resample is the proportion in $A$ of individuals belonging in the extent of $s$. Then, the hypergeometric law gives the probability of $X = x$ by:

$$\Pr(X=x) = \binom{n}{x} \binom{N-n}{N-n-x} / \binom{N}{n}$$

where $\binom{n}{x} = N! / n!(N-n)!$ is the number of possible samples of size $n$ in $N$, $\binom{N-n}{N-n-x} = (N-n)! / (n-x)!(N-n-x)!$ is the number of groups of $x$ individuals belonging in the extent of $s$ in $\Omega'$ and $\binom{N-n}{N-n-x} = (N-n)! / (n-x)!(N-n-x)!$ is the number of groups of $(n-x)$ individuals which are not belonging in the extent of $s$ in $\Omega'$. If the operator $T$ produces $k$ symbolic objects of extent in $A$ with size $x_1, \ldots, x_k$ then the more $Y = \sum_{i=1}^{k} \Pr(X = x_i)/k$ is small, the more these symbolic objects are characteristic of the class $A$. This happen for instance, when $p$ is small and $x/n$ large or $p$ large and $x/n$ small. Notice that in the case where $s$ is a complete symbolic object the size of the extent is $n$ and $p = n/N$, so

$$\Pr(X=n) = \binom{n}{n} \times \binom{0}{N-n} / \binom{N}{n} = 1 \times 1 / \binom{N}{n} = ((N-n)! / n!)/ N!$$

which is the probability of a complete symbolic object of size $n$ in a population of size $N$. When bootstrapping $\Omega'$, if the mean of the random variable $Y$ is out of the chosen confidence interval, then the more its standard deviation is low the more the characterisation is reliable. If we are interested by the variation of the characteristic of a specific symbolic objet, notice that at each resample we have to recognise each symbolic object. This can be done by the use of a dissimilarity measure between symbolic objects from one resample to the next (see for instance Bock, Diday (2000), chapter 8). The closest are considered to be the same.
the most influential individuals in a factorial axis, from a decision tree, etc.

Finally we can summarise the mathematical framework of a symbolic data analysis in the following way (figure 1):

Figure 1. \( \Omega \): set of individuals. \( D \): description set. \( L = \{\text{true, false}\} \) or \( L = [0,1] \). \( S \): set of symbolic objects. \( y \): description function. \( a \): membership function from \( \Omega \) in \( L = \{\text{true, false}\} \) 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. \( w_s = F(w) = (a, R, y(w)) \) 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

In figure 2 the "set of individuals" and the "set of concepts" is considered to be in the "real world", the "modeled world" is the "set of descriptions" which models individuals (or classes of individuals) and the "set of symbolic objects" which models concepts. We start with a "concept" \( C \) whose extent denoted \( \text{Ext}(C/\Omega') \) is known in a sample \( \Omega' \) of individuals. For instance, if the concept is "insurance companies", for instance, 30 insurance companies among a sample \( \Omega' \) of 1000 companies. Each individual \( w \) of the extent of \( C \) in \( \Omega' \) is described by using the mapping \( Y \) such that \( Y(w) \) describe the individual \( w \). The concept \( C \) is modelled 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 \( \text{Ext}(C/\Omega') \) 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).
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] and then the
symbolic object modelling the concept C is the triple s = (a_C, R, d_C).

When we don't have concepts as input, we get them in the following way:
i) A clustering of Ω' by using the description of the individuals produces a set of classes.

ii) To each interesting class denoted A, we associate a concept C and 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.

iii) The concept C is considered to be modelled by s_A.

**Quality, robustness and reliability of a symbolic object**

By definition of T, in the Boolean case, the extent of the symbolic object s_C contains the
extent of C but can contain also individuals which are not in the extent of C. 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 α: Ext_α( s_C) = {w/ a(w) ≥ α}. 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 don't satisfy the concept but are in the extent of s_C.

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

In order to validate the symbolic object s_C, we suggest the following bootstrap method:
Step1. Resample n times with replacement the initial sample Ω'.
Step2. Calculate the symbolic object s_C by following the scheme of figure 2 or figure 3.
Step3. Calculate the extent of s_C in Ω'.
Step4. Build the two histograms of the frequency of 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 is the lowest. In other words, let X1 (resp. X2) be the
random variable which associates to each resample 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 $s_C$. If the operator $T$ is the Min or the Max $t$-conorm, it is easy to show under a natural assumption that the smallest (resp. largest) symbolic object obtained by resampling converge in probability towards the symbolic object obtained by the Min (resp. Max) on the initial sample. The process has the following advantages:

- It reduces the importance of the outliers.
- It gives an idea on the errors distributions without any model assumption.
- It allows the comparison of different choices of the operator $T$ and the relation $R$.

The "reliability" of the membership of an individual $w$ to the extent of $s_C$ is measured by the mean $m(w)$ of the $a_C(w)$ after $n$ resample. If the $i$ th resample gives the value $a_i(w)$ then $m(w) = \sum_{i=1}^{n} a_i(w)/n$. and the reliability of $s_C$ is defined by:

$$W(s_C) = \sum_{w \in \text{Ext}(C) \cap \Omega} m(w)/|\text{Ext}(C) \cap \Omega|$$

The higher (i.e. the closest to 1) $W(s_C)$ is the better is the reliability of $s_C$.

The "sensibility" of $s_C$ to the bootstrap can be measured by:

$$W'(s_C) = \sum_{w \in \text{Ext}(C) \cap \Omega} \sigma(w)/|\text{Ext}(C) \cap \Omega|$$

with $\sigma(w)^2 = \sum_{i=1}^{n} (a_i(w)-m(w))^2/n$.

**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 to find outliers. In that case, we need to extend the notion of "mean" 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.
Figure 2. Modelisation by a symbolic object of a concept known by its extent
5) Some advantages in the use of symbolic objects

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

1. 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).

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

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

4. 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.

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

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

6) Some symbolic data analysis methods

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 all the advantages mentioned in 5).

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.

The following methods are developed in Bock, Diday (2000) and in the SODAS software:

- Principal Component and Discriminate Factorial Analysis of a symbolic data table. The
output of these 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.
- extension of elementary descriptive statistics to symbolic data (central object, histograms, dispersion, co-dispersion, etc. from a symbolic data table).
- extracting symbolic objects from the answers to queries of a relational data base,
- partitioning, hierarchical or pyramidal clustering of a set of individuals described by a symbolic data table such that each class be associated with a complete symbolic object.
- dissimilarities between Boolean or probabilistic symbolic objects,
- extension of decision trees on probabilistic symbolic objects, extension of a Parzen discrimination method to classes of symbolic objects,
- generalisation by a disjunction of symbolic objects of a class of individuals described in a standard way.
- inter-active and ergonomic graphical representation of symbolic objects.

7) Symbolic Data Analysis in the SODAS software

7.1 The general aim

The general aim of SODAS can be stated in the following way: building symbolic data in order to summarise huge data sets and then, 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:

<table>
<thead>
<tr>
<th>Household number</th>
<th>Region</th>
<th>Bedroom</th>
<th>Dining-Living</th>
<th>Socio-Econ group</th>
</tr>
</thead>
<tbody>
<tr>
<td>11404</td>
<td>Northern-Metropolitan</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11405</td>
<td>Northern-Metropolitan</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>11406</td>
<td>Northern-Metropolitan</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>12111</td>
<td>Northern-Metropolitan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12112</td>
<td>East anglia</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>12112</td>
<td>East anglia</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>12112</td>
<td>Greater London N-E</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

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 the table 5:

<table>
<thead>
<tr>
<th>Region</th>
<th>Bedroom</th>
<th>Dining-Liv</th>
<th>Socio-Ec gr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northern-Metropolitan</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Northern-Metropolitan</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Northern-Metropolitan</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Northern-Metropolitan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>East-anglia</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>East-anglia</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>East-anglia</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Greater London</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>North-East</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 the 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 variable 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.
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 data base (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 Data Bases Two 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,...).

7.2 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 data base. 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.

7.3 SODAS software overview

In figure 2 an overview on the SODAS software is given. The input of DB2SO (see chapter 5 in Bock, Diday (2000)) is a query to a data base. Its output is a symbolic data table. Having obtained this data table any symbolic data analysis method can be applied.

![Figure 2: A SODAS software overview](image)

7.4 SODAS future

The next steps in the future of SODAS will mainly consists 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 Data Base for instance, at different times (in order to study the time evolution of the retained concepts) or towards other data bases associated to different countries. In any case, we have to compare sets of concepts and their associated symbolic objects obtained from different data bases. 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 ways, 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 3.
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,...) to new kind of data table called "symbolic data table" and to give more explanatory results expressed by real world concepts mathematically represented by easy 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 was involved in this project: EUSTAT (Spain), INE (Portugal) and ONS (England). An example of future application 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 behaviour of the two countries. In that way, several new theoretical development are needed as the selection and the stochastic convergence of symbolic objects. Also, as the consensus between set of symbolic objects and their associated concepts extracted from different data bases. New software development are also needed as a tool in order to be able to transform a symbolic object extracted from a data base in a query of this data base or of another data base. 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 network etc.

References

Bertrand P. , Goupil F. (2000) "Descriptive statistics for Symbolic Data". In "Analysis of
Brito P. (1994) "Order structure of symbolic assertion objects". IEEE TR. on Knowledge and Data Engineering Vol.6, n° 5, October.
Diday E. et al. (1979) "Optimisation en classification automatique". INRIA edit. Rocquencourt 78150, France.

Roux M. (1985) "Algorithmes de classification", Masson..
Vignes (1991) "Caractérisation automatique de groupes biologiques". Thèse de doctorat, Université Paris 9 Dauphine.