Far Beyond the Classical Data Models: Symbolic Data Analysis

Monique Noirhomme-Fraiture[1] and Paula Brito[2]*

[1] *Faculté d'Informatique, Facultés Universitaires Notre-Dame de la Paix (FUNDP), Rue Grandgagnage, 21, B-5000 Namur, Belgium*

[2] *Faculdade de Economia & LIAAD/INESC Porto LA, Universidade do Porto, Rua Dr. Roberto Frias, 4200-464 Porto, Portugal*

Received 29 November 2010; revised 21 January 2011; accepted 22 January 2011
DOI:10.1002/sam.10112
Published online in Wiley Online Library (wileyonlinelibrary.com).

Abstract: This paper introduces symbolic data analysis, explaining how it extends the classical data models to take into account more complete and complex information. Several examples motivate the approach, before the modeling of variables assuming new types of realizations are formally presented. Some methods for the (multivariate) analysis of symbolic data are presented and discussed. This is however far from being exhaustive, given the present dynamic development of this new field of research. © 2011 Wiley Periodicals, Inc. Statistical Analysis and Data Mining, 2011

Keywords: data mining; symbolic data analysis; interval data; histogram data; imprecise data

1. INTRODUCTION

Since its introduction by E. Diday in the late eighties of the last century [1,2], symbolic data analysis (henceforth, SDA) has known a considerable development. It emerged from the need to consider data that contain information which cannot be represented within the classical data models, together with the objective of designing methods that produce results directly interpretable in terms of the input descriptive variables. The first European research project, "Symbolic Objects Data Analysis System", SODAS in short, gathered 17 teams working in SDA, including National Statistical Institutes (NSI's), in the years 1996–1999. The project allowed for a more systematic development of data analysis methodologies for symbolic data, producing the first statistical package for SDA—named SODAS—which made possible for data analysis researchers and users alike to produce, edit and analyze symbolic data. At the same time, the very first book on SDA was published, "Analysis of Symbolic Data" [3]. The SODAS project was then followed by a second European project, "Analysis System of Symbolic Official data", ASSO, gathering 15 teams, among which were three NSI's. The ASSO project gave the opportunity to continue the work of the SODAS project, develop new methodologies, update modules on the package and produce new ones, and publish a second book—"Symbolic Data Analysis and the Sodas Software" [4]. Official statistics appeared hence as a natural field of application for SDA methodologies, as studies in this area generally rely on aggregated—and therefore, more complex—data; also, confidentiality issues require usually the use of aggregated rather than individual data.

Since then, the development of SDA has continued at an intensive rate, with a growing number of publications in international journals, another book published [5], workshops and tutorials at the occasion of international conferences. The community of researchers in SDA includes nowadays teams in many different countries from all over the world, who are developing new methodologies and providing applications of SDA in a large number of different fields. In ref. 6, in this volume, a brief introduction to Symbolic Data and issues arising with their analyses are discussed, motivating the reader for the need to enlarge the scope of classical Data Analysis.

In this paper, we give an introduction to the field of SDA. Section 2 explains why the classical data model does not

Correspondence to: Paula Brito (mpbrito@fep.up.pt)

comply with a growing complexity of data, and motivates the introduction of a new approach. In Section 3, we present in more detail the different kinds of symbolic data, introducing the new types of realizations for the variables. In Section 4, we briefly discuss some of the problems that may arise when we are no longer in the framework of classical data. Section 5 gives a summary of some of the methods proposed for analyzing symbolic data. Finally, Section 6 concludes the paper, pointing out perspectives of future research.

2. FROM CLASSICAL TO SYMBOLIC DATA

In statistics and multivariate data analysis, the basic units under analysis are usually single individuals (often called "statistical units") which are described by a set of quantitative (i.e. numerical) and/or qualitative (also called categorical) variables, each individual taking one single value—which might possibly be missing or unknown—for each variable. For instance, a specific man may be described by his age, height, weight, his cranial perimeter, color of the eyes, etc. In describing a specific customer of a shop, one could inquire about his/her age, education, income, amount spent on that particular day, frequency of visit to the shop, etc. Data are often organized in a matrix or data array, where each cell (i, j) contains the value of variable j for individual i.

Table 1 presents an example of such a data array, where, for each person s_1 to s_4, two quantitative variables (number of children and weight) and two qualitative variables (gender and instruction level) are registered.

This model is however too restricted to take into account variability and/or uncertainty which are often inherent to the data. When analyzing a group rather than a single individual, then variability intrinsic to the group should be taken into account. Consider, for instance, that we are interested in analyzing the staff of some given teaching institutions, in terms of age, marital status and category. If we just take averages or mode values within each institution, much information is lost. Also, when we observe some given variables along time, and wish to record the set of observed values and not only a specific value (e.g. mean, maximum,...), then again a set of values rather than a single one must be recorded. The same issue arises when we are interested in concepts rather than in a single specimen—whether it is a plant species (and not the specific plant I have in my hand), a model of car (and not the particular one I am driving), etc.

Table 1. A classical data array.

	Number of children	Weight (kg)	Gender	Instruction level
s_1	2	52	M	2
s_2	1	55	M	3
s_3	0	50	M	2
s_4	3	60	F	1

Other examples may include a scientist specialist in fossils who defines the different species of dinosaurs—ceratosaurus, tyrannosaurus, iguanodons—after studying the skeleton of the animals and the environment where they were discovered; an economist could define different classes of stocks, based on the interest of the investors, e.g. high risk, medium risk, lower risk; a chemist could define different classes of wine based on chemical measurements. In each case, internal variation within each class is to be considered, providing aggregated generalized descriptions.

Whether the data are obtained by contemporaneous or temporal aggregation of individual observations to obtain descriptions of the entities which are of interest, or whether we are facing concepts as such specified by experts or put in evidence by clustering, we are dealing with elements which can no longer be properly described by the usual quantitative and qualitative variables without an unacceptable loss of information.

SDA provides a framework where the variability observed may effectively be considered in the data representation, and methods be developed that take it into account.

This approach is of particular and growing interest in the analysis of huge sets of data, recorded in very large databases, when the units of interest are not the individual records (the *microdata*), but rather some second-level entities. For instance, in a database of credit card purchases, we are probably more interested in describing the behavior of some person (or even some predefined class or group of persons) rather than each purchase by itself. By aggregating the purchase data for each person (or group), we obtain the information of interest; here again the observed variability for each person or within each group is of utmost importance.

To describe groups of individuals or concepts, variables may now assume other forms of realizations [3], which allow taking into account the intrinsic variability. These new variable types have been called "symbolic variables", and they may assume multiple, possibly weighted, values for each case. As in the classical statistics framework, we are dealing here with random variables, which may be observed in a given population; the term "symbolic" is used to stress the fact that the values they take are of a different nature.

As an example, consider the time needed for a person to go for work which varies from day to day, or the means of transportation used, which may be car, bus, etc. In the first case, the "value" for this variable is an interval

Table 2. Symbolic data table.

	Age	Marital status	Staff category
Institution 1	[20,45]	{single, married}	Administration (30%), Teaching (70%)
Institution 2	[30,50]	{married, divorced}	Administration (20%), Teaching (60%), Cleaning (20%)
Institution 3	[25,60]	{single, married, widow}	Administration (20%), Teaching (80%)

(e.g. [20 min, 40 min]), and in the second case, a frequency distribution (e.g. car 40%, bus 60%).

The information about the staff of some given teaching institutions could be presented as in Table 2.

As in the classical case, data are presented in a matrix, now called a "symbolic data table", each cell containing "symbolic data". Each row of the table corresponds to a group, or concept, i.e. the entity of interest; each row contains therefore its "symbolic description", each column corresponding to a "symbolic variable".

The process of obtaining aggregated information in the form of a symbolic data matrix from *microdata* in a relational database—called "DataBase to Symbolic Objects", DB2SO, in short—has been addressed within the framework of SODAS (see ref. 3), and improved subsequently (see ref. 4). It uses generalization operators, specific to variable type, and produces data matrices where the entities of interest are described by variables with new types of realizations (i.e. intervals, sets, histograms, etc.)

In the next sections, we define the different types of data that we meet in SDA. We start by giving the general definitions and set notation, and then discuss each variable type in more detail, illustrating with examples.

The SDA framework moreover allows taking into account some special structure that the variables may present. In this text, we will briefly discuss the case of taxonomic variables and hierarchical dependency between variables.

3. SYMBOLIC DATA: INTRODUCING NEW VARIABLE TYPES

Similarly, as in the classical case, we distinguish quantitative or qualitative symbolic variables or, in other terms, a symbolic variable may be *numerical* or *categorical*. This is a basic distinction, because the mathematical meaning and the operations that may be applied to either case are quite different for numerical and categorical variables.

Different kinds of numerical and categorical variables may then be considered; we will include classical quantitative and qualitative variables as special cases of symbolic variables.

A numerical (or quantitative) variable may then be *single-valued* (real or integer), as in the classical framework, if it takes one single value of an underlying domain per individual. It is *multi-valued* if its values are finite subsets of the domain and it is an *interval variable* if its values are intervals. When an empirical distribution over a set of subintervals is given, the variable is called a *histogram* variable. Other possibilities include variables whose values are a function, a time series or a symbolic stochastic process.

A categorical (or qualitative) variable can be *single-valued* (ordinal or not), as in the classical context, when it takes one category from a given finite domain $O = \{m_1, \ldots, m_k\}$ for each individual; *multi-valued*, if its values are finite subsets of the domain. A *categorical modal* variable Y with a finite domain $O = \{m_1, \ldots, m_k\}$ is a multistate variable where, for each element, we are given a category set and, for each category m_ℓ, a frequency or probability p_ℓ which indicates how frequent or likely that category is for this element.

Let $Y_1, \ldots, Y_p$ be the set of variables, O_j the underlying domain of Y_j and B_j the range of Y_j, $j = 1, \ldots, p$, i.e. the set where the variable takes its value for each entity.

A *description* is defined as a p-tuple $(d_1, \ldots, d_p)$ with $d_j \in B_j$, $j = 1, \ldots, p$. Let $S = \{s_1, \ldots, s_n\}$ be the entities, then $Y_j(s_i) \in B_j$ for $j = 1, \ldots, p, i = 1, \ldots, n$; therefore, the data array consists of n descriptions, one for each individual $s_i \in S : (Y_1(s_i), \ldots, Y_p(s_i)), i = 1, \ldots, n$.

Figure 1 below represents the ontology for the different variable types.

In the following text, we will denote Y a variable, O its underlying domain and B its range, i.e. the set where the variable takes its values for each entity.

3.1. Quantitative Variables

3.1.1. Quantitative single-valued variables

Given a set of n "individuals" $S = \{s_1, \ldots, s_n\}$, a quantitative single-valued variable Y is defined by an application

$$Y : S \to O \text{ such that } s_i \to Y(s_i) = \alpha$$

where $O \subseteq$ |R. That is, in this case, the range of Y, B, is identical to the underlying set O, $B \equiv O$.

This is the standard numerical variable. In SDA, we do not often meet such standard variables; the values being often issued from measures on many individuals, the probability to observe the same value in the *microdata* is usually zero. Exceptions may however occur for discrete variables.

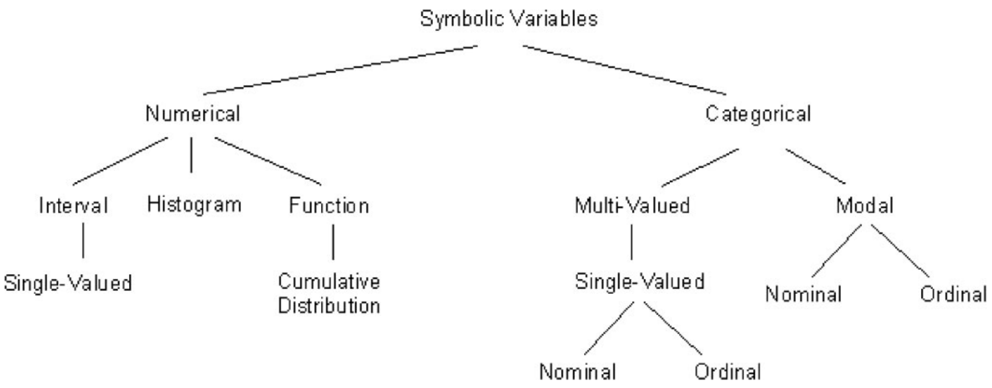

Fig. 1 Ontology of symbolic variables

3.1.2. *Quantitative multi-valued variables*

Given a set of n "individuals" S, a quantitative multi-valued variable Y is defined by an application

$$Y : S \to B \text{ such that } s_i \to Y(s_i) = \{\alpha_{i1}, \ldots, \alpha_{\mathrm{in}_i}\}$$

where B is the set of finite subsets of an underlying set $O \subseteq |\mathrm{R}$. The "values" of $Y(s_i)$ are hence finite sets of real numbers.

3.1.3. *Interval variables*

Given $\mathrm{S} = \{s_1, \ldots, s_n\}$, an interval variable is defined by an application

$$Y : S \to B \text{ such that } s_i \to Y(s_i) = [l_i, u_i]$$

where B is the set of intervals of an underlying set $O \subseteq |R$.

Let I be an $n \times p$ matrix representing the values of p interval variables on S. Each $s_i \in S$ is represented by a p-tuple of intervals, $I_i = (I_{i1}, \ldots, I_{ip})$, $i = 1, \ldots, n$, with $I_{ij} = [l_{ij}, u_{ij}]$, $j = 1, \ldots, p$ (see Table 3).

Interval data may occur in many different situations. We may have *native* interval data, when describing ranges of variable values—for example, daily stock prices or temperature ranges. Interval variables also allow dealing with imprecise data, coming from repeated measures or confidence interval estimation. A natural source of interval data is the aggregation of huge data bases, when real values describing the individual observations lead to intervals describing the aggregated data. Also, instead of the overall minimum and maximum values, percentiles, e.g. the tenth and the ninetieth percentile of the observed distribution may be considered, so as to exclude outliers. Original symbolic data—concerning, for instance, descriptions of biological species or technical specifications—constitute yet another possible source of interval data.

Table 3. Interval data table.

	Y_1	...	Y_j	...	Y_p
s_1	$[l_{11},u_{11}]$	...	$[l_{1j},u_{1j}]$	...	$[l_{1p},u_{1p}]$
...	...	...	...	...	...
s_i	$[l_{i1},u_{i1}]$	...	$[l_{ij},u_{ij}]$	...	$[l_{ip},u_{ip}]$
...	...	...	...	...	...
s_n	$[l_{n1},u_{n1}]$	...	$[l_{nj},u_{nj}]$	...	$[l_{np},u_{np}]$

The value of an interval variable Y_j for each $s_i \in S$ is hence defined by the bounds l_{ij} and u_{ij} of $I_{ij} = Y_j(s_i)$. For modeling purposes, however, an alternative parameterization consisting in representing $Y_j(s_i)$ by the midpoint $c_{ij} = (l_{ij} + u_{ij})/2$ and range $r_{ij} = u_{ij} - l_{ij}$ of I_{ij} may be useful.

Consider, as an example, a dataset containing information about patients (adults) in healthcare centers, during the second semester of 2008; Table 4 presents data for 3 of those centers. In healthcare center A, the age of patients ranged from 25 to 53 years old, each patient had 0, 1 or 2 emergency consults and the pulse measurement of the patients ranged from 44 to 86. Here, age and pulse are interval variables whereas the number of emergency consults is a multi-valued quantitative variable. The same description may be obtained for the remaining centers. Notice that in this example the entities under analysis are the healthcare centers, for each of which we have aggregated information, and not the individual patients in each center.

Table 4. Data for healthcare centers (1).

Healthcare center	Age Y_1	Number of emergency consults Y_2	Pulse Y_3
A	[25,53]	{0,1,2}	[44,86]
B	[33,68]	{1,4,5,10}	[54,76]
C	[20,75]	{0,5,7}	[70,86]

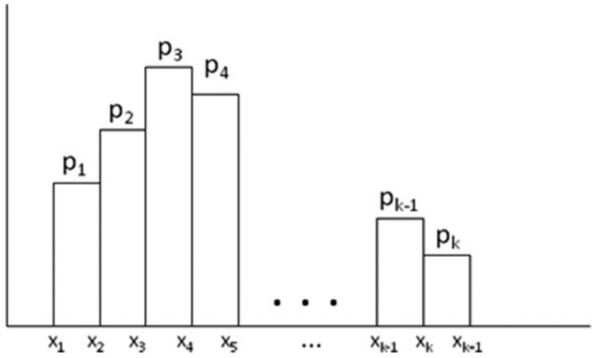

Fig. 2 Representation of the value of a histogram variable

3.1.4. Histogram variables

When real-valued *microdata* are aggregated by means of intervals, the information on the internal variation inside the intervals is lost: is the distribution Uniform, Normal or more special? One way to keep more information about this is to define limits between the global lower (LB) and upper bounds (UB) and compute frequencies between these limits. We obtain for each case and this numerical variable a histogram with k classes (and k frequencies) if $k-1$ is the number of limits between LB and UB.

Given S $= \{s_1, \ldots, s_n\}$, a histogram variable is defined by an application

$$Y : S \to B \text{ such that } s_i \to Y(s_i) = (I_{i1}(p_1), \ldots, I_{ik}(p_k))$$

where $I_{i1}, \ldots, I_{ik}$ are the considered subintervals, $p_{i1} + \ldots + p_{ik} = 1$; B is now the set of frequency distributions in $\{I_{i1}, \ldots, I_{ik}\}$ (Fig. 2).

It is generally assumed that, for each observation s_i, values are uniformly distributed within each subinterval. Also, for different observations, the number and length of subintervals of the histograms may naturally be different.

Consider again the healthcare centers example, with a new variable which records the waiting time for consultations. In this case, information is recorded for 5-time lengths (0–15 min, 15–30 min, etc.), and the corresponding variable is therefore a histogram variable (see Table 5).

Notice that the values of a histogram variable may equivalently be represented by an empirical distribution function:

$$F(x) = 0 \text{ for } x \le x_1$$
$$F(x) = p_1(x - x_1)/(x_2 - x_1) \text{ for } x_1 \le x \le x_2$$
$$F(x) = F(x_2) + p_2(x - x_2)/(x_3 - x_2) \text{ for } x_2 \le x \le x_3$$
$$F(x) = F(x_k) + p_k(x - x_k)/(x_{k+1} - x_k) \text{ for } x_k \le x \le x_{k+1}$$
$$F(x) = 1 \text{ for } x_{k+1} \le x$$

Also notice that when $k = 1$ a histogram reduces to an interval. Interval variables may hence be considered special cases of histogram variables.

3.1.5. More general quantitative variables

Apart from the above-described types of quantitative variables, more general cases may and have been considered. These include, for instance, functional variables, where for each entity a function is recorded. A special case of interest is that of cumulative percentage distributions, which has been investigated in refs 7,8.

Each concept could also be described by a time series, e.g. each stock is described by the evolution of its value against time.

More generally, we can consider that at each time, the variable is known by an interval: each day, the price of a stock is given by [min, max], and we can study the evolution of the interval from day to day. The study of such a process is rather complex; some hypothesis have to be considered like a Markovian one (see ref. 9).

3.2. Qualitative Variables

3.2.1. Categorical single-valued variables

This is the standard categorical variable. Given a set of n “individuals” $S = \{s_1, \ldots, s_n\}$, a qualitative single-valued variable is defined by an application

$$Y : S \to O \text{ such that } s_i \to Y(s_i) = m_\alpha$$

where O is a finite set of categories, $O = \{m_1, \ldots, m_k\}$ (i.e. in this case, again, $B \equiv O$).

Table 5. Data for healthcare centers (2).

Healthcare center	Age	Number of emergency consults	Pulse	Waiting time for consultation (min)
	Y_1	Y_2	Y_3	Y_4
A	[25,53]	{0,1,2}	[44,86]	([0,15[(0), [15,30[(0.25), [30,45[(0.5), [45,60[(0), ⩾ 60 (0.25))
B	[33,68]	{1,4,5,10}	[54,76]	([0,15[(0.25), [15,30[(0.25), [30,45[(0.25), [45,60[(0.25), ⩾ 60 (0))
C	[20,75]	{0,5,7}	[70,86]	([0,15[(0.33), [15,30[(0), [30,45[(0.33), [45,60[(0), ⩾ 60 (0.33))

In the symbolic context, it means that all the individuals of the concept share the same categorical value. For instance, all dinosaurs of the categories “ceratosaurus” and “tyrannosaurus” are carnivorous.

If the categories of O are naturally ordered, the variable is called *ordinal*, otherwise it is *nominal*.

Often, analysts use such a categorical variable to build new concepts or entities, by aggregating the cases sharing the same category.

3.2.2. Categorical multi-valued variables

A categorical multi-valued variable is defined by an application

$$Y : S \to B \text{ such that } s_i \to Y(s_i) = \{m_{i1}, \ldots, m_{\mathrm{in}_i}\}$$

where B is the set of finite subsets of an underlying set $O = \{m_1, \ldots, m_k\}$. The “values” of $Y(s_i)$ are hence finite sets of categories.

3.2.3. Categorical modal variables

A *categorical modal* variable Y with a finite domain $O = \{m_1, \ldots, m_k\}$ is a multistate variable where, for each element, we are given a category set and, for each category m_l, a weight, frequency or probability p_l which indicates how frequent or likely that category is for this element. If the weight is a frequency, it represents the proportion of individuals of the concept characterized by this category. Of course, the sum of the frequencies must add up to 1. That is, in this case, B is the set of distributions on O, and its elements are denoted $\{m_1(p_1), \ldots, m_k(p_k)\}$.

Consider again the healthcare centers example, and consider now that the education level of the patients has been recorded. We could then have a categorical modal variable Y_5, like in Table 6.

Another example may be considered coming from the field of official statistics. Suppose that answers to a survey on time use in some region were aggregated by gender and age, leading to concepts like F25/34 (female aged between 25 and 34 years old), F35/54, ... , M25/34. If frequencies are kept in the aggregation process, we can see for example that women between 35 and 54 years old (F35/54) feel hurried with the following distribution: frequently 27%, sometimes 22%, always 50%. We obtain thus the modal variable value {frequently (0.27); sometimes (0.22); always (0.50)}; by contrast, for F25/34, the following “value” has been obtained {frequently (0.36); sometimes (0.26); always (0.38)}.

The weights could moreover be something else rather than probabilities or frequencies, such as capacities [10], necessities, possibilities and credibilities [11,12]. In these cases, the sum of the different weights does not necessarily add up to 1.

Categorical modal variables are similar to histogram variables for the quantitative case, in that their values are both characterized by classes (or categories) and weights. Nevertheless, from a mathematical point of view, they are quite different and the vocabulary should not be mixed.

Notice that a multi-valued categorical variable could be considered as a particular case of a modal one: in the multi-valued case, if we set all frequencies equal for the categories, we obtain a modal one. This corresponds however to the assumption of the uniformity of the distribution, which is not the same as ignoring the frequencies.

3.3. Taxonomic Variables

Let $Y : S \to O$. The variable Y is a taxonomic variable if O is ordered into a tree structure. Taxonomies may be taken into account in obtaining descriptions of aggregated *microdata*: first, values are recorded as in the case of categorical multi-valued variables and then each set of

Table 6. Data for healthcare centers (3).

Healthcare center	Age Y_1	Number of emergency consults Y_2	Pulse Y_3	Waiting time for consultation (min) Y_4	Education level Y_5

A	[25,53]	{0,1,2}	[44,86]	([0,15[(0), [15,30[(0.25), [30,45[(0.5), [45,60[(0), $\geq$ 60 (0.25))	{9th grade, 1/2;
					Higher education, 1/2}

B	[33,68]	{1,4,5,10}	[54,76]	([0,15[(0.25), [15,30 [(0.25), [30,45[(0.25), [45,60[(0.25), $\geq$ 60 (0))	{6th grade, 1/4;
					9th grade, 1/4;
					12th grade, 1/4;
					Higher education, 1/4}

C	[20,75]	{0,5,7}	[70,86]	([0,15[(0.33), [15,30[(0), [30,45[(0.33), [45,60 [(0), $\geq$ 60 (0.33))	{4th grade, 1/3;
					9th grade, 1/3;
					12th grade, 1/3}

values of O is replaced by the lowest value in the taxonomy covering all the values of the given set. We choose to go up to level h when at least two successors of h are present:

$$[\text{Form} = \{\text{triangle, rectangle}\}] \rightarrow [\text{Form} = \{\text{polygon}\}].$$

3.4. Hierarchical Rules

A variable Y' is said to be hierarchically dependent from a variable Y if it cannot be applied if Y takes values within a given set $A : Y$ takes values in A $\Leftrightarrow Y'$ is not applicable. In other words, we say that a variable Y' is *hierarchically dependent* on a variable Y if Y' makes no sense for some values of Y, and hence becomes "non-applicable". For instance, if in a survey we want to describe the previous job of a person, the variable does not apply when the present job is the first one. Descriptions that do not comply with a rule are called "non-coherent". For instance, the description (Yes, secretary), associated with [First job = {Yes}] $\wedge$ [Previous job = {secretary}] is non-coherent. The consideration of hierarchical rules in SDA has been widely studied in refs 13–15.

4. PROBLEMS ARISING IN THE ANALYSIS OF SYMBOLIC DATA

The need to consider data that go beyond the classical model, where each "individual" takes exactly one value per variable has lead to the development of SDA. To represent the data taking into account internal variability within each observation, variables have been allowed to assume new forms. However, are we still in the same framework when we allow for the variables to take multiple values? Are the definitions of basic statistical notions still so straightforward? What properties remain valid? These are some of the issues that arise when it is wished to apply classical statistical and multivariate data analysis techniques to symbolic data.

Let us consider the case of quantitative variables. Here, the central question of the evaluation of dispersion, and the consequences of different possible choices in the design of multivariate methods, has to be addressed. Dispersion is, for instance, a key issue in clustering, since the result of any clustering method depends heavily on the scales used for the variables. The standardization problem has been addressed in ref. 16, and different standardization techniques for interval-type variables have been proposed. Furthermore, many exploratory multivariate methodologies rely heavily on the notion of linear combinations and on the properties of dispersion measures under linear transformations. How should a linear combination of symbolic quantitative variables be defined? This problem has been addressed in a recent work [17] in the context of a linear discriminant analysis of interval data.

Different approaches have been considered by various authors to address these and other questions and propose a symbolic counterpart of statistical multivariate data analysis methods. Most existing methods for the analysis of such data rely however on a nonparametric descriptive approach. The important issue in analyzing symbolic data remains the need for models; without statistical modeling, no estimation or hypothesis testing is possible.

In the next section, we make a non-exhaustive overview of methods that have been designed and proposed to analyze different kinds of symbolic data.

5. METHODS FOR THE ANALYSIS OF SYMBOLIC DATA

In this section, we make a short description of some of the approaches and methods proposed to analyze symbolic data.

It should be noticed however that there has been a considerable greater effort in addressing and designing methods for interval data rather than for any other type of symbolic data. Furthermore, some methodologies have received more attention than others.

5.1. Univariate and Bivariate Descriptive Statistics

For the case of interval variables an equidistribution hypothesis is assumed in ref. 18, which consists in considering each observation as equally likely and that the values of the underlying variable are uniformly distributed. This approach assumes that each interval variable represents the possible values of an underlying real-valued variable. The empirical distribution function of an interval variable is then defined as a uniform mixture of n uniform distributions. More specifically, we have, for every $x \in |R$,

$$F_j(x) = \frac{1}{n} \sum_{i=1}^{n} \Pr(X_{ij} \leq x)$$

where X_{ij} is a uniformly distributed random variable in the interval $[l_{ij}, u_{ij}]$.

From this assumption, it was then straightforward to derive expressions for the empirical mean and variance of interval variables.

Following the same reasoning, the joint density function of two interval variables has been derived in ref. 19. Later on, in refs 20,21, the empirical covariance has been derived by studying how the total, between interval and within interval variations relate. A different approach, using copulas, may be found in ref. 22.

In refs [19,21], this approach is also extended to the histogram data case, and in ref. 23 to interval-valued observations in the presence of rules.

Statistical descriptive measures for numerical and categorical multi-valued variables are obtained in ref. [18], also in the presence of rules. The case of modal multi-valued variables is considered in ref. 5.

5.2. Factorial Analysis

Principal component analysis (PCA) of interval data has first been addressed in refs 24,25, either by representing the observed intervals by their centers ("centers method") or by considering all the vertices of the hypercube representing each of the n individuals in a p-dimensional space ("vertices method"). A different approach is based on representing each variable by the midpoints and ranges of its interval values, a line of work followed in ref. 26. Three methods for PCA of fuzzy interval data are discussed in ref. 27.

The extension of PCA to histogram variables has been addressed in refs 28,29, by representing each observation described by histogram variables by a succession of k interval nested entities (k being the maximum number of subintervals).

Instead of representing the histograms in the factorial plane, they represent the empirical distribution function associated with each histogram as defined in ref. 3. More recently, using a novel approach, Ichino [30] proposes a method for symbolic PCA for histogram variables, based on a quantile representation of the data.

Generalized canonical analysis, allowing for the factorial analysis of different variable types (interval, categorical multi-valued, modal), using a multistep symbolic-numerical-symbolic procedure is developed in ref. 31.

5.3. Clustering Approaches

Clustering is a multivariate methodology that aims at organizing similar entities in homogeneous classes, on the basis of the observed values in a set of variables. The classes may be organized according to different structures. Hierarchical and pyramidal clustering methods produce a structure of nested clusters; in the case of a hierarchy each level corresponds to a partition (i.e. by "cutting" a hierarchy at an appropriate level, we obtain a partition of the data set S); in the case of a pyramid, we find, at each level, a family of overlapping clusters (family of nonempty subsets of S which together cover S but are not necessarily disjoint), but all clusters are intervals of a total linear order. Partitional (nonhierarchical) clustering methods produce directly, by means of an iterative process, a partition of S on a generally predefined number of disjoint classes, by—most generally locally—optimizing some given criteria.

To analyze the data that go beyond the classical models it was necessary to define, or adapt, clustering methods. Moreover, it was intended that the clusters found should be represented within the same formalism as the input data, since symbolic variables allow describing classes, taking into account their internal variability [1,2].

Since the initial formalization of symbolic data and the first steps in SDA [1,2], a multitude of methods for clustering symbolic data has been proposed and studied, and applied in different domains. We categorize these methods into two distinct groups [32]:

(1) Methods that result from adapting classical clustering methods based on dissimilarities to the new kind of data, by properly defining dissimilarity measures for symbolic data. In this case, the clustering methodologies and criteria remain almost unchanged (only necessary adaptations have to be performed) and are applied to the obtained dissimilarity matrices.

(2) Methods that do not rely on dissimilarities and use the data (i.e. the descriptions of the elements of S) explicitly in the clustering process. The criterion to form classes is to obtain a "meaningful" class description, and we are in the scope of the so-called conceptual clustering methods.

It should be noticed that this categorization is not specific to the clustering of symbolic data, the same applies in the case of clustering classical data arrays. Clustering methods of type (1) will tend to cluster together entities with similar descriptions—this similarity being evaluated by one of the proposed measures—irrespective of the intrinsic variability of the underlying descriptions. In other words, however large is the variability inherent to two given descriptions, if they are alike, their dissimilarity will have a low value—and the corresponding entities will tend to be clustered together. On the other hand, methods of type (2) will tend to concentrate on the description of each newly formed cluster, and minimize its inherent variability. This means that this kind of method may favor the grouping of entities whose descriptions are less alike, if the description of the resulting cluster presents a lower variability.

This duality is specific to symbolic data; it does not arise if we are in the presence of classical—quantitative or qualitative—data. In the latter case, the closer the values of a given variable, the more specific is their generalization—so both dissimilarity and generalization based methods will tend to elect the same candidate pairs to be aggregated. The criteria are different and therefore it makes no sense to compare results issued by the two kinds of methods, since they start from a different concept of “what a cluster is”.

When clustering is based on dissimilarities and the underlying variables are quantitative, then the question of comparability of the measurement scales of the different variables is a major issue. It is well known, and may often be verified in practical applications, that the dissimilarity values and, consequently, the clustering results are strongly affected by the variables’ scales; so, to make it possible to obtain an objective or scale-invariant result, some standardization must be performed prior to dissimilarity computations in the clustering process.

In ref. 16, three alternative standardization methods for the case of interval data have been proposed; these methods mainly differ in the way dispersion of an interval-valued variable is evaluated. An alternative approach for the standardization for interval data was proposed in ref. [33], when a Hausdorff distance is used. In that paper, the author points out that to compute distances between standardized observations is generally equivalent to using a normalized version of the corresponding distance.

Several dissimilarity measures for different types of symbolic data, adopting different points of view on how to measure dissimilarity in this new context, have been proposed and investigated. For an overview and discussion on different alternatives, refer to ref. 34 or 35.

Many methods based on different dissimilarities, generally adaptations of *k-means* have been developed, see refs 16,33,36–40. Fuzzy approaches have been considered in ref. 41. Other extensions, using adaptive distances (i.e. distances that vary from cluster to cluster) have also been proposed, see refs 42–46.

A method based on Poisson point processes has been proposed in ref. 47. Clustering and validation of interval data are discussed in ref. 48.

A method for “symbolic” hierarchical or pyramidal clustering has been proposed in refs [49–51], which allows clustering multi-valued data of different types. This method was subsequently developed in order to allow for modal variables [52]; later on, this was extended so as to allow for the existence of hierarchical rules between multi-valued categorical variables [53] and between modal variables [54]. The method may be seen within the framework of conceptual clustering, since each cluster formed is associated with a conjunction of properties in the input variables, which constitutes a necessary and sufficient condition for cluster membership. Clusters are hence associated with concepts, described extensionally by the set of its members and intentionally by a symbolic description expressing the variability of each variable within the cluster. An additional criterion has been considered to choose among the different aggregations meeting the above condition. The principle is that clusters associated with less general descriptions should be formed first. A measure has been defined that allows quantifying the generality of a given description: this is the so-called “generality degree”. For interval-valued and categorical multi-valued variables, this evaluates the proportion of the underlying domain that is covered by the symbolic description; for modal variables, it evaluates how much the given distribution is close to the uniform distribution, by computing the affinity between the given distribution and the uniform distribution (see ref. 55). The generality degree is computed variable-wise; the values for each variable are then combined in a multiplicative way to give a measure of the variability of the symbolic description.

In this context, mention should also be made of the monothetic clustering method, which uses a divisive approach, proposed in ref. 56. This applies both to interval and modal categorical variables, and uses a criterion that measures intra-class inertia.

An approach using Kohonen maps has been developed for clustering interval data by Bock [57].

On a recent approach, Irpino and Verde [58] propose using the Wasserstein distance for clustering histogram data.

More recently, in ref. 59, a new method for symbolic clustering based on quantile representation is presented.

5.4. Discriminant or Unsupervised Learning

In ref. 60, a generalization of classical factorial discriminant analysis to symbolic data is proposed. The method is based on a numerical analysis of transformed symbolic data, followed by a symbolic interpretation of the results; it allows considering quantitative, qualitative nominal or modal variables. Classification rules are then based on proximities in the factorial plane. See also ref. 61 for a more up-to-date version of the methodology.

Different approaches to discriminant analysis of interval data are compared in ref. 17. Three fundamental approaches are considered. The first approach assumes a uniform distribution in each observed interval, and appropriately defines linear combinations of interval variables that maximize the usual discriminant criterion. The second approach expands the original data set into the set of all interval description vertices, and proceeds with a classical analysis of the expanded set. Finally, a third approach replaces each interval by a midpoint and range representation. Resulting representations, using intervals or single points, are discussed and distance based allocation rules are proposed.

Symbolic kernel discriminant analysis, developed in ref. 62, is based on intensity estimation. Since the object space is nonintegrable, we cannot talk about densities in this context. Bayesian decision trees for the case when predictors are interval variables are presented in ref. 63.

A method for building a decision tree on probabilistic data is developed in ref. 64, the method may also apply to non-probabilistic (interval or multi-valued) variables, by transforming them to modal variables with uniform probability functions; for this methodology see also ref. 65.

A generalized recursive tree-growing method that includes a strata partition in the tree-building algorithm is presented in ref. 66. The objective is to predict and explain the value of the class variable by the predictors, conditioned to the stratum, explain how these predictions are affected by stratum membership, and detect sets of strata with similar explanation. An example, given in ref. 66, concerns a survey made both to employed and unemployed people, from different economic sectors (for the unemployed, the sector of the last job was recorded). The objective was to explain unemployment in the population by variables such as gender, age, education, etc., and have this related to the economic sector. In this example, the class variable to be explained is employed/unemployed, the predictors are the gender, age, education, etc. and the economic sectors are the different strata considered. That is, it is wished to obtain an explanation for the employment situation conditional on the economic sector, identify economic sectors for which this explanation is the same and describe each economic sector by the rules that may be applied for employment in this sector. The method applies to classical data as input as well as to categorical modal data.

Discriminant analysis of interval data has also been addressed using support vector machines [67,68], as well as artificial neural networks [69–72].

5.5. Linear Regression Models

Linear regression in SDA has mainly been addressed for the case of interval variables although some steps have also been done for histogram variables.

Using the empirical covariance obtained from the uniform approach, Billard and Diday [73] proposed the first linear regression model for interval-valued variables, which is in fact equivalent to a classical linear regression on the centers of the observed intervals; the estimated coefficients are then applied to the lower and upper bounds of the independent variables to estimate lower and upper bounds of the dependent variable. Notice, however, that some coefficients may be negative, so that these bounds must be chosen taking into account which of the obtained values is the lowest.

Along the same lines, later Billard and Diday [74] presented a series of models, among which the minmax model which, however, has the same characteristics of the centers model.

Using a different approach, Neto and De Carvalho [75] propose a new model, estimating the midpoint and half-range of the dependent variable from separate classical linear regressions on the independent variables' intervals' midpoints and half-ranges. This model requires two classical regressions, one for the midpoints and another one for the half-ranges. Again, there is no guarantee that the obtained half-range is positive, which should naturally be the case. Two years later [76], the same authors proposed a new model with nonnegativity restrictions on the mid-ranges regression coefficients, therefore imposing a direct relation between the ranges of the dependent and independent variables.

An extension of their first model in ref. 73 uses the obtained covariance values for the estimation of the regression coefficients, to the histogram data case [74].

5.6. Time Series Analysis

The first approach work [77] to consider interval-valued time series data used an approach based on fitting univariate autoregressive integrated moving average (ARIMA) processes to the interval bounds. Fitting univariate ARIMA processes applied to the midpoint and radius and using them to forecast the interval bounds was proposed in ref. 78; they also proposed an approach based on an artificial neural network model as well as a combination of both.

In refs 79–83, interval stochastic processes, interval-valued time series and weak stationarity for interval processes are defined and, based on the sample moments previously proposed [5,15], they also define the empirical autocovariance and autocorrelation functions for interval time series data, aiming at uncovering the data generating process behind this type of symbolic data sets. In refs 79–81, the focus is on forecasting based on vector autoregressive models (VAR models) and vector error correction models (VEC) and smoothing filters. In ref. 80, VAR and interval multilayer perceptrons are compared for electric power demand forecasting using interval time series.

Also, the problem of forecasting for histogram variables is addressed in refs. [79,84]. The mean error between the observed and forecast histogram values is based on Mallows or Wasserstein distances, which the authors consider most adequate for this purpose.

5.7. Galois Lattices

Following the definition of Galois connections [85], the Galois lattice associated with an object-variable correspondence raised interest in binary data analysis, which has been particularly emphasized by Barbut and Monjardet [86]. Thereafter, the importance of the Galois lattice of a relation has been widely recognized. Many theoretical and algorithmic developments have been accomplished, both by, for instance, Wille [87], Ganter and Wille [88], and by Duquenne and Guigues [89]. In these studies, lattice theory is used for the analysis, organization and interpretation of the data. Galois lattices allow identifying structured clusters, which are automatically interpreted, putting in evidence the links among the objects, among the variables, and between objects and variables.

The extension of Galois connections and Galois lattices to symbolic data has first been addressed by Brito [49–51], by developing the appropriate operators for the new variable types; this has then been further developed in refs 90–92. In ref. 93, two Galois connections on a set of probabilistic data are defined, and the corresponding concept lattices, and the notion of "complete symbolic object" is extended to these data; a new algorithm for the construction of the concept lattice is also presented; applications illustrate the interest of the approach.

6. CONCLUSION AND PERSPECTIVES

SDA extends statistics and multivariate data analysis to more complex data, by providing a framework which allows taking into account variability and/or uncertainty inherent to the data. Variables are now allowed to assume new types of realizations, which appropriately model these data, therefore avoiding unnecessary loss of information. In recent years, different approaches have been investigated and many methods proposed for the analysis of symbolic data. These have however not developed uniformly across data types and analysis methods: interval data is by far the most investigated and for which more methods have been developed; cluster analysis has received considerably more attention from researchers than any other multivariate methodology. Applications in different domains have proved the well-founding and usefulness of the proposed approaches.

Much remains however to be done. We have just rather briefly discussed some of the issues that arise when we leave the classical data framework and allow for more complex variable types. Usual properties, generally taken for granted, often do not apply any longer, and new concepts much be put forward. Among these, parametric statistical analysis is an important challenge. The first promising steps in this direction have been made in refs 94–96. Anyhow, a whole world of problems still remains open, waiting to be explored.

ACKNOWLEDGMENTS

The authors would like to thank the guest editor for the valuable suggestions that helped to improve the manuscript. The authors also would like to thank Dr. Sónia Dias, from the University of Porto, for kindly providing the data on healthcare centers, used in Section 3.

REFERENCES

[1] E. Diday, The symbolic approach in clustering and related methods of data analysis: the basic choices, In Classification and Related Methods of Data Analysis, Proceedings of IFCS'87, H.-H. Bock, ed., Aachen, July 1987, North Holland, Amsterdam, 1988, 673–684.

[2] E. Diday, Introduction à l'approche symbolique en analyse des données, RAIRO, Recherche Opérationnelle 23(2), (1989), 193–236.

[3] H.-H. Bock and E. Diday, eds., Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data, Berlin-Heidelberg, Springer-Verlag, 2000.

[4] E. Diday and M. Noirhomme-Fraiture, eds., Symbolic Data Analysis and the Sodas Software, Chichester, Wiley, 2008.

[5] L. Billard and E. Diday, Symbolic Data Analysis: Conceptual Statistics and Data Mining, Chichester, Wiley, 2006.

[6] L. Billard, Brief overview of symbolic data and analytic issues, Stat Anal Data Min, this issue, (2011).

[7] E. Diday and M. Vrac, Mixture decomposition of distributions by copulas in the symbolic data analysis framework, Discrete Appl Math 147(1) (2005), 27–41.

[8] E. Cuvelier, QAMML: Probability Distributions for Functional Data. Ph.D. Thesis, University of Namur, Belgium, 2009.

[9] M. Noirhomme-Fraiture, Asymptotic behaviour in symbolic Markov chain, In Classification as a Tool for Research, In Proceedings of the 11th IFCS Conference, Dresden, H. Locarek-Junge, C. Weihs, eds., Heidelberg, Springer, 2010.

[10] G. Choquet, Theory of capacities, Anna Inst Fourier, 5 (1954), 131–295.

[11] D. Dubois and H. Prade, Properties of measures of information in evidence and possibility theories, Fuzzy Sets and Systems 100 Supplement, 1999, 35–49.

[12] P. Walley, Towards a unified theory of imprecise probability, International Journal of Approximate Reasoning 24(2–3) (2000), 125–148.

[13] R. Vignes, Caractérisation Automatique de Groupes Biologiques, Ph.D. Thesis, University Paris VI, 1991.

[14] F. A. T. De Carvalho, Proximity coefficients between boolean symbolic objects, In New Approaches in Classification and Data Analysis, E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy, eds., Berlin-Heidelberg, Springer-Verlag, 1994, 387–394.

[15] M. Csernel and F. A. T. De Carvalho, Usual operations with symbolic data under Normal Symbolic Form, Applied Stochastic Models in Business and Industry 15 (1999), 241–257.

[16] F. A. T. De Carvalho, P. Brito, and H.-H. Bock, Dynamic clustering for interval data based on L_2 distance, Comput Stat 21(2) (2006), 231–250.

[17] A. P. Duarte Silva and P. Brito, Linear discriminant analysis for interval data, Comput Stat 21(2) (2006), 289–308.

[18] P. Bertrand and F. Goupil, Descriptive statistics for symbolic data, In Analysis of Symbolic Data, Exploratory Methods

for Extracting Statistical Information from Complex Data, H.-H. Bock and E. Diday, eds., Heidelberg, Springer, 2000, 106–124.

[19] L. Billard and E. Diday, From the statistics of data to the statistics of knowledge: symbolic data analysis, J Am Stat Assoc 98(462) (2003), 470–487.

[20] L. Billard, Dependencies and variation components of symbolic interval-valued data, In Selected Contributions in Data Analysis and Classification, P. Brito, P. Bertrand, C. Cucumel, and F. De Carvalho, eds., Heidelberg, Springer, 2007, 3–12.

[21] L. Billard, Sample covariance functions for complex quantitative data, In Proceedings of IASC2008, Joint Meeting of 4th World Conference of the IASC and 6th Conference of the Asian Regional Section of the IASC on Computational Statistics & Data Analysis, Yokohama, Japan, 2008.

[22] L. Billard, Dependencies in bivariate interval-valued symbolic data, In Classification, Clustering and Data Mining Applications, D. Banks, L. House, F. R. McMorris, P. Arabie, W. Gaul, eds., Proceedings of the Meeting of the International Federation of Classification Societies (IFCS 2004), Berlin-Heidelberg, Springer, 2004, 319–324.

[23] L. Billard and E. Diday, Descriptive statistics for interval-valued observations in the presence of rules, Comput Stat 21(2) (2006), 187–210.

[24] A. Chouakria, P. Cazes, and E. Diday, Symbolic principal component analysis, In Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data, H.-H. Bock and E. Diday, eds., Heidelberg, Springer, 2000, 200–212.

[25] P. Cazes, A. Chouakria, E. Diday, and Y. Schektman, Extensions de l'analyse en composantes principales à des données de type intervalle, Rev Stat Appl 24 (1997), 5–24.

[26] C. Lauro and F. Palumbo, Principal component analysis for non-precise data, In New Developments in Classification and Data Analysis, M. Vichi, P. Monari, S. Mignani, and A. Montanari, eds., Berlin-Heidelberg, Springer, 2005, 173–184.

[27] P. Giordani and H. A. L. Kiers, A comparison of three methods for principal component analysis of fuzzy interval data, Comput Stat & Data Anal, special issue The Fuzzy Approach to Statistical Analysis 51(1) (2006), 379–397.

[28] O. Rodriguez, E. Diday, and S. Winsberg, Generalization of the principal components analysis to histogram data, In Proceedings 4th European Conference on Principles and Practice of Knowledge Discovery in Data Bases; Workshop on Symbolic Data Analysis, Lyon, 14, 2000.

[29] O. Rodriguez and A. Pacheco, Applications of histogram principal components analysis, In The 15th European Conference on Machine Learning (ECML) and the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Pisa, 2004.

[30] M. Ichino, Symbolic PCA for histogram-valued data, In Proceedings of IASC2008, Joint Meeting of 4th World Conference of the IASC and 6th Conference of the Asian Regional Section of the IASC on Computational Statistics & Data Analysis, Yokohama, Japan, 2008.

[31] C. Lauro, R. Verde, and A. Irpino, Generalized canonical analysis, In Symbolic Data Analysis and the Sodas Software, E. Diday and M. Noirhomme-Fraiture, eds., Chichester, Wiley, 2008, 313–330.

[32] P. Brito, On the analysis of symbolic data, In: Selected Contributions in Data Analysis and Classification, P. Brito, P. Bertrand, C. Cucumel, and F. De Carvalho, eds., Heidelberg, Springer, 2007, 13–22.

[33] M. Chavent, Normalized k-means clustering of hyper-rectangles, In Proceedings of the XIth International Symposium of Applied Stochastic Models and Data Analysis (ASMDA 2005), Brest, France, 2005, 670–677.

[34] F. Esposito, D. Malerba, and A. Appice, Dissimilarity and matching, In Symbolic Data Analysis and the Sodas Software, E. Diday and M. Noirhomme-Fraiture, eds., Chichester, Wiley, 2008, 123–148.

[35] E. Diday and F. Esposito, An introduction to symbolic data analysis and the SODAS software, Intelligent Data Analysis 7 (2003), 583–602.

[36] R. M. C. R. de Souza and F. A. T. De Carvalho, Clustering of interval data based on City-Block distances, Pattern Recogn Lett 25(3) (2004), 353–365.

[37] R. M. C. R. de Souza, F. A. T. De Carvalho, and C. P. Tenorio, Two partitional methods for interval-valued data using Mahalanobis distances, IBERAMIA, 2004, 454–463.

[38] M. Chavent, F. A. T. De Carvalho, Y. Lechevallier, and R. Verde, New clustering methods for interval data, Comput Stat 21(2) (2006), 211–229.

[39] F. A. T. De Carvalho and R. M. C. R. de Souza, Unsupervised pattern recognition models for mixed feature-type symbolic data, Pattern Recogn Lett 31(5) (2010), 430–443.

[40] F. A. T. De Carvalho, M. Csernel, and Y. Lechevallier, Clustering constrained symbolic data, Pattern Recogn Lett 30(11) (2009), 1037–1045.

[41] F. A. T. De Carvalho, Fuzzy c-means clustering methods for symbolic interval data, Pattern Recogn Lett 28 (2007), 423–437.

[42] R. M. C. R. de Souza, F. A. T. De Carvalho, and F. C. D. Silva, Clustering of interval-valued data using adaptive squared Euclidean distances, In Proceedings ICONIP, 2004, 775–780.

[43] R. M. C. R. de Souza and F. A. T. De Carvalho, Dynamic clustering of interval data based on adaptive Chebyshev distances, Electron Lett 40(11) (2004), 658–659.

[44] F. A. T. De Carvalho, R. M. C. R. de Souza, M. Chavent, and Y. Lechevallier, Adaptive Hausdorff distances and dynamic clustering of symbolic interval data, Pattern Recogn Lett 27(3) (2006), 167–179.

[45] F. A. T. De Carvalho and Y. Lechevallier, Partitional clustering algorithms for symbolic interval data based on single adaptive distances, Pattern Recogn 42(7) (2009), 1223–1236.

[46] F. A. T. De Carvalho and C. P. Tenorio, Fuzzy k-means clustering algorithms for interval-valued data based on adaptive quadratic distances, Fuzzy Sets and Systems 161(23) (2010), 2978–2999.

[47] A. Hardy and N. Kasaro, A new clustering method for interval data, Mathématiques et Sciences Humaines 187 (2009), 79–91.

[48] A. Hardy and J. Baune, Clustering and validation of interval data, In Selected Contributions in Data Analysis and Classification, P. Brito, P. Bertrand, C. Cucumel, and F. De Carvalho, eds., Heidelberg, Springer, 2007, 69–82.

[49] P. Brito, Analyse de Données Symboliques. Pyramides d'Héritage. Ph.D. Thesis, University Paris-IX Dauphine, 1991.

[50] P. Brito, Use of pyramids in Symbolic Data Analysis, In New Approaches in Classification and Data Analysis, E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and

B. Burtschy, eds., Berlin-Heidelberg, Springer-Verlag, 1994, 378–386.

[51] P. Brito, Symbolic objects: order structure and pyramidal clustering, Ann Oper Res 55 (1995), 277–297.

[52] P. Brito, Symbolic clustering of probabilistic data, In Advances in Data Science and Classification, A. Rizzi, M. Vichi, and H.-H. Bock, eds., Berlin-Heidelberg, Springer-Verlag, 1998, 385–390.

[53] P. Brito and F. A. T. De Carvalho, Symbolic clustering in the presence of hierarchical rules, In Studies and Research Proceedings of the Conference on Knowledge Extraction and Symbolic Data Analysis (KESDA'98), Luxembourg, Office for Official Publications of the European Communities, 1999, 119–128.

[54] P. Brito and F. A. T. De Carvalho, Symbolic clustering of constrained probabilistic data, In Exploratory Data Analysis in Empirical Research, O. Opitz, M. Schwaiger, eds., Heidelberg, Springer Verlag, 2002, 12–21.

[55] P. Brito and F.A.T. De Carvalho, Hierarchical and pyramidal clustering, In Symbolic Data Analysis and the Sodas Software, E. Diday and M. Noirhomme-Fraiture, eds., Chichester, Wiley, 2008, 181–203.

[56] M. Chavent, A monothetic clustering method, Pattern Recognition Letters 19(11) (1998), 989–996.

[57] H.-H. Bock, Visualizing symbolic data by Kohonen maps, In Symbolic Data Analysis and the Sodas Software, E. Diday and M. Noirhomme-Fraiture, eds., Chichester, Wiley, 2008, 205–234.

[58] A. Irpino and R. Verde, A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data, Data Science and Classification, Proceedings of the Conference of the International Federation of Classification Societies (IFCS06), Berlin, Springer, 2006, 185–192.

[59] P. Brito and M. Ichino, Symbolic clustering based on quantile representation, presented at COMPSTAT'2110, Paris, 2010.

[60] N. C. Lauro, R. Verde, and F. Palumbo, Factorial discriminant analysis on symbolic objects, In Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data, H.-H. Bock and E. Diday, eds., Heidelberg, Springer, 2000, 212–233.

[61] N. C. Lauro, R. Verde, and A. Irpino, Factorial discriminant analysis, In Symbolic Data Analysis and the Sodas Software, E. Diday and M. Noirhomme-Fraiture, eds., Chichester, Wiley, 2008, 341–358.

[62] J. P. Rasson and S. Lissoir, Symbolic kernel discriminant analysis, In Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data, H. -H. Bock and E. Diday, eds., Heidelberg, Springer, 2000, 240–244.

[63] J. P. Rasson, J.-Y. Pirçon, P. Lallemand, and S. Adans, Unsupervised divisive classification, In Symbolic Data Analysis and the Sodas Software, E. Diday and M. Noirhomme-Fraiture, eds., Chichester, Wiley, 2008, 149–156.

[64] E. Périnel and Y. Lechevallier, Symbolic discrimination rules, In Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data, H.-H. Bock and E. Diday, eds., Heidelberg, Springer, 2000, 244–265.

[65] A. Ciampi, E. Diday, J. Lebbe, E. Périnel, and R. Vignes, Growing a tree classifier with imprecise data, Pattern Recogn Lett 21(9) (2000), 787–803.

[66] M. C. Bravo Llatas and J. M. Santesmases, Segmentation Trees for Stratified Data, In Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data, H.-H. Bock and E. Diday, eds., Heidelberg, Springer, 2000, 266–293.

[67] T.-N Do and F. Poulet, Kernel methods and visualization for interval data mining, In Proceedings of the Conference on Applied Stochastic Models and Data Analysis, ASMDA 2005. J. Janssen and P. Lenca, eds., ENST Bretagne, 2005.

[68] E. Carrizosa, J. Gordillo, and F. Plastria, Classification problems with imprecise data through separating hyperplanes' [Online]. Available at http://www.optimization-online.org/DB_FILE/2007/09/1781.pdf, 2007.

[69] J. Síma, Neural expert systems, Neural Netw 8(2) (1995), 261–271.

[70] S. J. Simoff, Handling uncertainty in neural networks: an interval approach', In Proceedings of the IEEE International Conference on Neural Networks, IEEE, Washington DC, 1996, 606–610.

[71] M. Beheshti, A. Berrached, A. de Korvin, C. Hu, and O. Sirisaengtaksin, On interval weighted freelayer neural networks, In Proceedings of the 31st Annual Simulation Symposium, IEEE Computer Society Press, 1998, 188–194.

[72] F. Rossi and B. Conan Guez, Multilayer perceptron on interval data, In Classification, Clustering and Data Analysis, K. Jajuga, A. Sokolowski, and H.-H. Bock, eds., Berlin, Heidelberg, New York, Springer, 2002, 427–434.

[73] L. Billard and E. Diday, Regression analysis for interval-valued data, in 'Data Analysis, Classification, and Related Methods, In Proceedings of the Seventh Conference of the International Federation of Classification Societies (IFCS00), Springer, 2000, 369–374.

[74] L. Billard and E. Diday, Symbolic regression analysis, Classification, Clustering and Data Analysis, In Proceedings of the Conference of the International Federation of Classification Societies (IFCS02), Springer, 281–288, 2002.

[75] E. A. L. Neto and F. A. T. De Carvalho, Centre and range method for fitting a linear regression model to symbolic interval data, Comput Stat Data Anal 52(3) (2008), 1500–1515.

[76] E. A. L. Neto and F. A. T. De Carvalho, Constrained linear regression models for symbolic interval-valued variables, Computational Statistics & Data Analysis 54(2) (2010), 333–347.

[77] P. Teles and P. Brito, Modelling interval time series data, In Proceedings of the 3rd IASC World Conference on Computational Statistics and Data Analysis, Limassol, Cyprus, 2005.

[78] A. L. S. Maia, F. A. T. De Carvalho, and T. D. Ludermir, Forecasting models for interval-valued time series, Neurocomputing 71(16–18) (2008), 3344–3352.

[79] J. Arroyo, Métodos de Predicción para Series Temporales de Intervalos e Histogramas. Ph.D. Thesis, Universidad Pontifícia Comillas, Madrid, Spain, 2008.

[80] C. García-Ascanio and C. Maté, Electric power demand forecasting using interval time series: a comparison between VAR and iMLP, Energy Policy 38 (2009), 715–725.

[81] J. Arroyo, G. González-Rivera, and C. Maté, Forecasting with interval and histogram data. Some financial applications, In Handbook of Empirical Economics and Finance, A. Ullah, D. Giles, N. Balakrishnan, W. Schucany, and E. Schilling, eds., Chapman and Hall/CRC, New York, 2010.

[82] G. González-Rivera and J. Arroyo, Time series modeling of histogram-valued data: The daily histogram time series of S&P500 intradaily returns, Int. J. Forecasting (in press).

[83] A. Han, Y. Hong, K. Lai, and S. Wang, Interval time series analysis with an application to the Sterling-Dollar exchange rate, J Syst Sci Complex 21(4) (2008), 558–573.

[84] J. Arroyo and C. Maté, Forecasting histogram time series with k-nearest neighbours methods, Int J Forecast 25 (2009), 182–207.

[85] G. Birkoff, Lattice Theory, Vol. XXV (3rd ed.). American Mathematical Society Colloquium Publications, Providence, 1967.

[86] M. Barbut and B. Monjardet, Ordre et Classification, Algèbre et Combinatoire, Tomes I et II, Hachette, Paris, 1970.

[87] R. Wille, Restructuring lattice theory: an approach based on hierarchies of concepts, In Proceedings of the Symposium on Ordered Sets, I. Rival, ed., Dordrecht-Boston, Reidel, 1982, 445–470.

[88] B. Ganter and R. Wille, Formal Concept Analysis—Mathematical Foundations, Berlin, Springer Verlag, 1999.

[89] V. Duquenne and J. L. Guigues, Familles minimales d'implication informatives résultant d'un tableau de données binaires, Math Sci Hum 95 (1986), 5–18.

[90] G. Polaillon, Organisation et Interprétation par les Treillis de Galois de Données de Type Multivaluée, Interval ou Histogramme, Ph.D. Thesis, Université Paris IX Dauphine, 1998.

[91] G. Polaillon, Interpretation and reduction of Galois lattices of complex data, In Advances in 'Data Science and Classification', A. Rizzi, M. Vichi, and H.-H. Bock, eds., Springer-Verlag, Berlin, 1998, 433–440.

[92] G. Polaillon and E. Diday, Reduction of symbolic Galois lattices via hierarchies, In Proceedings of the Conference on Knowledge Extraction and Symbolic Data Analysis (KESDA'98), Office for Official Publications of the European Communities, Luxembourg, 1999, 137–143.

[93] P. Brito and G. Polaillon, Structuring probabilistic data by Galois lattices, Mathématiques et Sciences Humaines - Mathematics and Social Sciences, (43ème année) nb.169, (1), 2005, 77–104.

[94] H.-H. Bock, Probabilistic modeling for symbolic data, In COMPSTAT - Proceedings in Computational Statistics, P. Brito, ed., Heidelberg, Springer, 2008, 55–65.

[95] P. Brito and A. P. Duarte Silva, Modeling interval-data with Normal and Skew-Normal distributions, In Proceedings of IASC2008, Joint Meeting of 4th World Conference of the IASC and 6th Conference of the Asian Regional Section of the IASC on Computational Statistics & Data Analysis, Yokohama, Japan, 2008.

[96] J. Le-Rademacher and L. Billard, Likelihood functions and some maximum likelihood estimators for symbolic data, Journal of Statistical Planning and Inference 141 (2011), 1593–1602.