Overview

# Symbolic Data Analysis: another look at the interaction of Data Mining and Statistics

Paula Brito*

Symbolic Data Analysis (SDA) provides a framework for the representation and analysis of data that comprehends inherent variability. While in Data Mining and classical Statistics the data to be analyzed usually presents one single value for each variable, that is no longer the case when the entities under analysis are not single elements, but groups gathered on the basis of some given criteria. Then, for each variable, variability inherent to each group should be taken into account. Also, when analysing concepts, such as botanic species, disease descriptions, car models, and so on, data entail intrinsic variability, which should be explicitly considered. To this purpose, new variable types have been introduced, whose realizations are not single real values or categories, but sets, intervals, or, more generally, distributions over a given domain. SDA provides methods for the (multivariate) analysis of such data, where the variability expressed in the data representation is taken into account, using various approaches. 

**How to cite this article:**
*WIREs Data Mining Knowl Discov* 2014, 4:281–295. doi: 10.1002/widm.1133

## INTRODUCTION

In Data Mining, Multivariate Data Analysis, and Classical Statistics, the data to be analyzed is usually represented in a $n \times p$ data array where each row represents an entity (a 'case' or an 'individual'), each column pertains to a variable (also called 'attribute'), which may be numerical or categorical, and one single value is recorded for each variable and for each of the $n$ entities. This representation model is however somehow restricted when the data to be analyzed entail variability. This is the case when the entities under analysis are not single elements, but rather groups formed on the basis of some given common properties. Then, for each descriptive variable, the observed variability inherent to each group should be taken into account to avoid a too important loss of pertinent information. This is also the case when analysing concepts as such—a botanical species, and not a given specimen; a car model, and not a specific vehicle, and so on—then again variability is intrinsic to the data and ought to be explicitly considered. As an example, suppose that a study on airport activity is being pursued, for which data are collected for each flight arriving at different airports, e.g., number of passengers, delay of arrival, aircraft company(ies), and so forth. As the statistical units of interest are the airports and not each individual flight, data concerning flights arriving on a same airport should be somehow aggregated. Another such situation arises when a study is performed for comparing different schools, and data are collected for individual students—age, gender, marks at different exams, and so on. Again, data of individual students must be aggregated for the corresponding school, so that descriptions of each school may be obtained and subsequently analyzed and compared. In situations of such kind, where the statistical units of interest are at a higher level of that at which data have been collected, aggregation of the observed values must be carried out prior to data analysis. The standard approach is to compute summary indicators—such as means, medians, or mode values—so that data fit into the usual $n \times p$ data array

*Correspondence to: mpbrito@fep.up.pt
Faculdade de Economia & LIAAD-INESC TEC, Universidade do Porto, Porto, Portugal
Conflict of interest: The author has declared no conflicts of interest for this article.

  

model, and classical methods may be applied. This procedure, however, obviously entails an important loss of information.

Symbolic data analysis (SDA; see, e.g., Ref 1) provides a framework for the representation and analysis of data with inherent variability. To this purpose, new variable types have been introduced, whose realizations are not single real values or categories, but sets, intervals, or, more generally, distributions over a given domain. Naturally, the analysis of such data raises new issues, because most concepts and methods have primarily been designed for single-valued observations. Until now, nevertheless, many methods for the (multivariate) analysis of symbolic data have been developed, following different approaches and using distinct criteria, which allow taking the variability expressed in the data representation into account.

Another approach where data are treated at aggregated level is Granular Computing—see, e.g., Ref 2. Information granules are defined as groups of individual observations, which capture the semantics of the abstract entities of interest in the problem at hand. Generally, given a dataset $D$, granulation results in a set of granules, formed on the basis of similarity or closeness, which may be accomplished, e.g., by clustering algorithms. When the data at hand are numerical, granules often take the form of hyper-rectangles. Information granules are then generally represented within the theory of fuzzy sets, i.e., by means of a membership function.

Surely, there are common points of view between the two approaches, in pursuing the aggregation of the original data into more general entities and proceeding with an analysis of elements at an higher order. Nevertheless, clear differences may be pointed out. In SDA, the original individual and the aggregated data are represented within a same framework of a $n \times p$ data array, where new variable types describe the formed groups, expressing explicitly their within variability and with no link to fuzzy set representations. As a direct consequence, the analysis methods designed for symbolic data, although possibly aiming at the same objectives (e.g., clustering, classification, etc.), rely on different measures and properties.

This being said, the two approaches do deserve a comparative study with some depth, both at theoretical and applied levels, so as to highlight their respective advantages and complementarities. This extends however beyond the scope of this study.

In the next section, we present and motivate the emergence of SDA into some more detail, and discuss different sources of symbolic data. In section *Variable Types*, we formalize the different variable types. Section *From Classical to Symbolic Data* discusses issues raised in the statistical analysis of symbolic data. Methods for the multivariate analysis of different types of symbolic data are recalled and overviewed in section *Methods for the Analysis of Symbolic Data*. Section *Conclusion* concludes the study with a general overview, and points out directions of current and future research.

## SYMBOLIC DATA

Symbolic data, i.e., data which contain intrinsic variability, arise in distinct multiple contexts. The most common is, however, from the aggregation of individual observations—the so-called microdata.

We may distinguish two different types of aggregation:

- Temporal aggregation: when data are recorded at different time moments for the same entities—e.g., subsequent purchases made by the same individuals at a given store—but time is not an issue (i.e., we are not interested in the chronological order of observations). Records must then be aggregated so that the whole set of values (or their distribution) is considered, and not just a mean, median, or mode value is kept. In this situation, observations are made along time, and the statistical units under analysis (the 'cases', i.e., the customers in the given example) are still the same before and after aggregation.
- Contemporary aggregation: this concerns the situation when data are recorded at the same point in time, but we are interested in analysing entities at a higher level than that at which data were originally collected, e.g., when data are collected for students and we are interested in comparing classes or schools as a whole. In this case, the statistical units that will be analyzed are not those for which data were originally recorded, but constitute specific groups of those. Notice that this case also comprises the situation of spatial aggregation, when data are collected at a same point in time for statistical units—e.g., individual citizens, as in official statistics surveys—across different regions, and then data are aggregated at region level, for a chosen granularity, such as parish, district, and so forth, so that the regions become the units of interest.

Let us consider examples of both situations.

As an illustration of temporal aggregation, consider three people, Albert, Barbara, and Caroline, who are characterized by the amount of time (in min) they

  

**TABLE 1** | Interval Data: Time to Go to Work

| Person | Time (min) |
|---|---|
| Albert | [15,20] |
| Barbara | [25,30] |
| Caroline | [10,20] |

need to go to work. This varies from day to day, and the observed variation may be expressed by intervals as in Table 1.

Other such situations are, for instance:

- Blood pressure repeatedly measured for different patients;
- Some technical quantity for which measures are taken at different points of given structures.

In all these cases, the statistical units are the same before and after data aggregation.

Now suppose, as mentioned above, that a study is being conducted on different schools, and that data are collected for students attending these schools, e.g., age, gender, marks at different exams, and so on. The statistical units of interest are the schools and not the individual students, and therefore data concerning students attending a same school must be aggregated. Table 2 represents an example of data collected for some students, and Table 3 the same data, aggregated by school.

**TABLE 2** | Student's Data

| Student's Name | School | Age | Gender | Math's Mark |
|---|---|---|---|---|
| Albert | A | 12 | M | 12 |
| Bertha | B | 11 | F | 14 |
| Charles | A | 12 | M | 15 |
| Deborah | A | 13 | F | 16 |
| Emily | B | 13 | F | 13 |
| Felicity | B | 12 | F | 11 |
| George | A | 14 | M | 10 |
| … | … | … | … | … |

**TABLE 3** | Data for Schools

| School | Age | Gender | Math's Mark |
|---|---|---|---|
| A | [12,14] | {*F*, *M*} | {< 10, (0.2); [10 – 15], (0.6); > 15, (0.2)} |
| B | [11,13] | {*F*} | {< 10, (0.3); [10 – 15], (0.3); > 15, (0.4)} |

In this small example, data for variable *Age* have been aggregated as intervals, for *Gender* in the form of sets, and for *Math's marks* as distributions.

Other examples of similar situations may be mentioned: when data are collected for individual players, but the units to be studied and compared are the teams as a whole; when data are collected for individual citizens (as in official statistics surveys), but studies are to be made on parishes, cities, of special sociographic groups of interest, and so forth.

This is particularly interesting in Data Mining applications where huge sets of data are collected, and data should be analyzed at a higher level. Consider the case of super markets or large department stores, which record data on each purchase made, e.g., the amount spent, items purchased, quantity of each item, and so on. Generally, administrations and marketing researchers are not particularly interested in each purchase by itself, but rather on consumer behavior. That is, they wish to have information about the overall purchases of each client, or specific groups of clients. To obtain such information, data collected for the individual purchases must be aggregated. Other such examples, addressed in Data Mining studies, concern: data about individual phone calls, which the communications' operator wishes to analyze aggregated at the client level; data about web logs, to be aggregated by user or website; and data about medical prescriptions—and the units of interest for the study are the doctors and/or the patients. The SDA framework provides the possibility to aggregate the individual 'microdata' keeping variability across records.

In recent years, the term 'Big Data' emerged, referring to data sets so large and complex that they become difficult to process with traditional data analysis applications and in a reasonable amount of time. SDA, offering the possibility of aggregating data at the user's chosen degree of granularity while keeping the information on the intrinsic variability, and then analyze the resulting (symbolic) data arrays, may play an important role in this context. Steps in this direction appear to being taken, see e.g., Ref 3 for the aggregation of data streams.

Giordano and Brito[4] use SDA for the study and comparison of social networks. In this work, a network symbolic description is defined according to the statistical characterization of the network topological properties and suitable network measures are represented by their distributions for each network. Multidimensional data analysis then allows for the synthetic representation of a network as a point onto a metric space and subsequent analysis (e.g., clustering).

SDA may also be of interest in (large) survey analysis, when the observed sample is partitioned into specific groups which are the focus of interest. This may be the case, e.g., in sociological or marketing surveys, when groups defined e.g., by age, gender, education level, and/or professional status should be compared and analyzed together. Moreover, SDA makes it possible to merge independent surveys made on a same population, at a macro level. In this case, the 'microdata' may not be analyzed together, as the observed individuals in the different surveys are not the same. By aggregating the concerned surveys using the same criteria (i.e., forming the same 'groups'), we obtain data which may be compiled together, by just juxtaposing the columns referring to the (symbolic) variables formed in each case.

## VARIABLE TYPES

To represent data variability, new variable types have been introduced in SDA, whose realizations are now not restricted to real values (in the numerical case) or individual categories (in the qualitative case). The different considered variable types, including the classical ones—which may be considered special cases of the symbolic types (see Ref 1)—are defined below.

As in classical Statistics, we distinguish numerical and categorical variables. A numerical (or quantitative) variable is single valued (real or integer), as in the classical framework, if it takes one single value of an underlying domain for each entity. It is multi-valued if its values are finite subsets of the domain and is an interval-valued variable if its values are intervals of $\mathbb{R}$. When a distribution over a set of subintervals is given, the variable is called a histogram-valued variable. A categorical (or qualitative) variable is single-valued (ordinal or nominal), when it takes one category from a given finite category set $O=\{m_1, \ldots, m_k\}$ for each entity; multi-valued, if its values are finite subsets of $O$. A categorical modal variable $Y$ with a finite domain $O=\{m_1, \ldots, m_k\}$ is a multi-valued variable, where for each element we are given a category set and, for each category $m_\ell$, a frequency or probability which indicates how frequent or likely that category is for this element.[5] Let $Y_1, \ldots, Y_p$ be the set of variables, $O_j$ the underlying domain of $Y_j$, and $B_j$ the set where $Y_j$ takes its value for each entity, for $j=1, \ldots, p$. A description $d$ is defined as a $p$-tuple $d=(d_1, \ldots, d_p)$ with $d_j \in B_j$, $j=1, \ldots, p$. Let $S=\{s_1, \ldots, s_n\}$ be the set of entities (the statistical units) under analysis, then $Y_j(s_i) \in B_j$ for $j=1, \ldots, p$, $i=1, \ldots, n$. The data array to be analyzed consists of $n$ descriptions, one for each $s_i \in S: d_i=(Y_1(s_i), \ldots, Y_p(s_i))$, $i=1, \ldots, n$.

### Classical Variables

#### *Quantitative Single-Valued Variables*

Given the set of $n$ entities $S=\{s_1, \ldots, s_n\}$, a quantitative single-valued variable $Y$ is defined by an application $Y: S \to O$ such that $s_i \mapsto Y(s_i)=c \in O \subseteq \mathbb{R}$. This is the classical numerical case, and $B$ is identical to the underlying set $O$, $B \equiv O$.

#### *Categorical Single-Valued Variables*

A categorical single-valued variable is a standard categorical variable. Given $S=\{s_1, \ldots, s_n\}$ and a finite set of categories, $O=\{m_1, \ldots, m_k\}$ a categorical single-valued variable is defined by an application $Y: S \to O$ such that $s_i \mapsto Y(s_i)=m_\ell$ (i.e., in this case, again $B \equiv O$). If the categories of $O$ are naturally ordered, the variable is called ordinal, otherwise it is nominal. Such a categorical variable may be used to build new concepts or entities by aggregating the cases sharing the same category.

### New Variable Types

#### *Quantitative Multi-Valued Variables*

Given the set $S$, a quantitative multi-valued variable $Y$ is defined by an application $Y: S \to B$ such that $s_i \mapsto Y(s_i)=\{c_{i1}, \ldots, c_{in_i}\}$. Here $B$ is the power set of an underlying set $O \subseteq \mathbb{R}$ (excepting the empty set $\emptyset$). $Y(s_i)$ is now a finite nonempty set of real numbers.

#### *Interval-Valued Variables*

Given $S=\{s_1, \ldots, s_n\}$, an interval-valued variable is defined by an application $Y: S \to B$ such that $s_i \mapsto Y(s_i)=[l_i, u_i]$, $B$ is the set of intervals of an underlying set $O \subseteq \mathbb{R}$. Let $I$ be an $n \times p$ matrix representing the values of $p$ interval-valued 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 4).

The value of an interval-valued variable $Y_j$ for each $s_i \in S$ is usually defined by the lower and upper bounds $l_{ij}$ and $u_{ij}$ of $I_{ij}=Y_j(s_i)$. For modeling purposes, however, it may be useful to represent $Y_j(s_i)$ by the midpoint $c_{ij}=(l_{ij}+u_{ij})/2$ and range $r_{ij}=u_{ij}-l_{ij}$ of $I_{ij}$.

**TABLE 4** | Matrix $I$ of Interval Data

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

  

**TABLE 5** | Data for Airports (1)

| Airport | No. Passengers | No. Companies |
|---|---|---|
| A | [150,200] | {1, 2} |
| B | [180,300] | {1, 2, 3} |
| C | [200,400] | {1, 3} |

*Example:* Consider a dataset containing information about arriving flights at some airports; Table 5 presents data of three airports. In airport A, for instance, the number of passengers in arriving flights ranges from 150 to 200, and the number of companies involved is 1 or 2. Here, the number of passengers is an interval variable, whereas the number of companies involved is a multi-valued quantitative variable. A similar description may be obtained for the remaining airports. It should be stressed that in this example the entities under analysis are the airports, for each of which we have aggregated information and NOT the individual flights.

In the study by Brito and Duarte Silva,[6] parametric models for interval data are proposed, which consider Multivariate Normal or Skew-Normal distributions for the MidPoints and Log-Ranges of the interval-valued variables. The Gaussian model has the advantage of allowing for the application of classical inference methods, the Skew-Normal setup allows for some more flexibility. In either case, it is important to keep in mind that the MidPoint $c_{ij}$ and the Range $r_{ij}$ of the value of an interval-valued variable $I_{ij} = Y_j(s_i)$ are two quantities related to one same variable and must be considered together. Therefore, the global covariance matrix should take into account the link that may exist between MidPoints and Ranges of the same or different variables. Intermediate parameterizations between the nonrestricted and the noncorrelation setup considered for real-valued data are relevant for the specific case of interval data. The following cases are of particular interest and have been addressed:

1. Nonrestricted case: allowing for nonzero correlations among all MidPoints and Log-Ranges;
2. Interval-valued variables $Y_j$ are uncorrelated, but for each variable, the MidPoint may be correlated with its Log-Range;
3. MidPoints (Log-Ranges) of different variables may be correlated, but no correlation between MidPoints and Log-Ranges is allowed;
4. All MidPoints and Log-Ranges are uncorrelated, both among themselves and between each other.

The referred modeling, for the Gaussian case, has been implemented in the R-package MAINT.Data,[7] available on CRAN. MAINT.Data introduces a data class for representing interval data and includes functions for modeling and analysing these data. In particular, maximum likelihood estimation and statistical tests for the considered configurations are addressed. Methods for (M)ANOVA and Linear and Quadratic Discriminant Analysis of this data class are also provided.

### *Histogram-Valued Variables*

When real-valued data are aggregated by means of intervals, the information on the distribution inside the intervals is not taken into account. One way to keep more detailed information is to define subintervals between the global lower (LB) and upper (UB) bounds and compute frequencies for these intervals. We obtain for each case a histogram with $k$ classes (and $k$ frequencies) where $k$ is the number of the considered subintervals. Naturally, to aggregate numerical microdata by means of a histogram implies that a reasonably large number of observations are available at the micro level. Given $S = \{s_1, \ldots, s_n\}$, a histogram-valued variable is defined by an application $Y : S \to B$ such that $s_i \mapsto Y(s_i) = \left\{ \left[\underline{I}_{i1}, \overline{I}_{i1}\right], p_{i1}; \left[\underline{I}_{i2}, \overline{I}_{i2}\right], p_{i2}; \ldots; \left[\underline{I}_{ik_i}, \overline{I}_{ik_i}\right], p_{ik_i} \right\}$ where $I_{i\ell} = \left[\underline{I}_{i\ell}, \overline{I}_{i\ell}\right]$, $\ell = 1, \ldots, k_i$ are the subintervals considered for observation $s_i$, $p_{i1} + \cdots + p_{ik_i} = 1$; $B$ is now the set of frequency distributions in $\left\{I_{i1}, \ldots, I_{ik_i}\right\}$. It is assumed that for each entity $s_i$ values are uniformly distributed within each subinterval. For different observations, the number and length of subintervals of the histograms may naturally be different.

*Example:* Consider again the airports example, with a new variable which records the delay (in min) of each arriving flight. In this case, information is recorded for three time lengths (0 to 10 min, 10 to 30 min, 30 min to 1 h), the corresponding variable is therefore a histogram-valued variable (see Table 6).

**TABLE 6** | Data for Airports (2)

| Airport | No. Passengers | No. Companies | Delay (min) |
|---|---|---|---|
| A | [150,200] | {1, 2} | {[0, 10[, 0.25; [10, 30[, 0.65; [30, 60[, 0.10} |
| B | [180,300] | {1, 2, 3} | {[0, 10[, 0.45; [10, 30[, 0.30; [30, 60[, 0.25} |
| C | [200,400] | {1, 3} | {[0, 10[, 0.75; [10, 30[, 0.20; [30, 60[, 0.05} |

The values of a histogram-valued variable may equivalently be represented by the empirical distribution function F or by its inverse, the quantile function Ψ. This latter option is often used, given that all quantile functions are defined in the same domain [0,1], which is convenient for comparisons. The quantile function associated with a histogram-valued observation $Y(s_i) = \left\{\left[\underline{I}_{i1}, \overline{I}_{i1}\right], p_{i1}; \left[\underline{I}_{i2}, \overline{I}_{i2}\right], p_{i2}; \ldots; \left[\underline{I}_{ik_i}, \overline{I}_{ik_i}\right], p_{ik_i}\right\}$ is given by:

$$\Psi(t) = F^{-1}(t) = \begin{cases} \underline{I}_{i1} + \frac{t}{w_{i1}} r_{I_{i1}} & \text{if} \quad 0 \le t < w_{i1} \\ \underline{I}_{I_{i2}} + \frac{t-w_{i1}}{w_{i2}-w_{i1}} r_{i2} & \text{if} \quad w_{i1} \le t < w_{i2} \\ \vdots \\ \underline{I}_{ik_i} + \frac{t-w_{ik_i-1}}{1-w_{ik_i-1}} r_{ik_i} & \text{if} \quad w_{ik_i-1} \le t \le 1 \end{cases}$$

where $w_{il} = 0$ if $l = 0$ and $w_{il} = \sum_{\ell=1}^{l} p_{i\ell}$ if $l = 1, \ldots, k_i$ and $r_{i\ell} = \overline{I}_{i\ell} - \underline{I}_{i\ell}$ with $\ell \in \{1, \ldots, k_i\}$; $k_i$ is the number of subintervals in $Y(s_i)$.

When $k = 1$ a histogram reduces to an interval: interval-valued variables may therefore be considered special cases of histogram-valued variables.

### *Categorical Multi-Valued Variables*

A categorical multi-valued variable is defined by an application $Y: S \to B$ where $B$ is the set of nonempty subsets of $O = \{m_1, \ldots, m_k\}$. The 'values' of $Y(s_i)$ are now finite sets of categories.

### *Categorical modal variables*

A categorical modal variable $Y$ with a finite domain $O = \{m_1, \ldots, m_k\}$ is a multi-valued variable where, for each element, we are given a category set and, for each category $m_\ell$, a weight, frequency, or probability $p_\ell$ which indicates how frequent or likely that category is for this element. It may be imposed that the sum adds up to 1, although this is not compulsory from the definition. In this case, $B$ is the set of distributions (probability, frequency, or other) on $O$, and its elements are denoted $\{m_1(p_1), \ldots, m_k(p_k)\}$.

*Example:* Consider again the airports example and the information on the main airline companies. We then have a categorical modal variable as shown in Table 7.

In fact, the weights may be something else rather than probabilities or frequencies, such as capacities,[8] necessities, possibilities, or credibilities.[9,10] In these cases, their sum does not necessarily add up to 1. For a more general discussion of these cases, see Ref 11 where the author shows how probabilist, possibilist, and belief theories may be extended to the analysis of symbolic data.

**TABLE 7** | Data for Airports (3)

| Airport | Main Companies |
|---|---|
| A | {British (0.25), Lufthansa (0.4), Air France (0.35)} |
| B | {British (0.10), Lufthansa (0.15), Air France (0.6), Iberia (0.15)} |
| C | {Lufthansa (0.3), Air France (0.5), Iberia (0.2)} |

*Example:* Suppose that in the previous example it is wished to generalize the description of airports A and B from the aggregated descriptions (the original microdata are no longer available). If we generalize by taking the maximum of the weights corresponding to each category, we obtain

$$\text{A\&B} : \{\text{British } (0.25), \text{ Lufthansa } (0.4), \text{ Air France } (0.6), \text{ Iberia } (0.15)\}$$

which should be interpreted as 'in airports A and B, at most 25% of the flights are from British Airways, 40% are Lufthansa flights, 60% are from Air France, and 15% are Iberia flights'. In this case, the set of the two airports are described by a possibilitic distribution on the airline categories.

Generalization of descriptions involving histogram-valued or categorical modal variables by the maximum or the minimum operators leads to distributions where the sum of the values for each class/category may not equal one. Those have been used in conceptual clustering of symbolic data, in e.g., Ref 12, see also Ref 13.

Categorical modal variables are similar to histogram-valued variables for the quantitative case, in that their values are both characterized by classes or categories and weights. Henceforth in this text, by 'distributional data' we refer to both types, as opposed to 'set-valued' variables, when no distribution is given. Nevertheless, from a mathematical point of view, they are of different nature.

## Quantile Representation

Quantile representation[14,15] provides a common framework to represent symbolic data described by variables of different types. It is based on the fact that a monotone property of symbolic descriptions is characterized by the nesting structure of the Cartesian join regions. On a discrete approach, the principle is to express the observed variable values by some predefined quantiles of the underlying distribution; however, variable values may be represented by the quantile function of the underlying distribution, therefore considering a continuous setup. For

  

**TABLE 8** | Symbolic Data Array for Airport Data

| Airport | Number Passengers | Delay Time on Arrival (min) | Distance Category |
|---|---|---|---|
| A | [150,200] | {[0, 10[, 0.25; [10, 30[, 0.65; [30, 60], 0.10} | {1 (0.40) ; 2 (0.40) ; 3 (0.2)} |
| B | [180,300] | {[0, 10[, 0.45; [10, 30[, 0.30; [30, 60], 0.25} | {1 (0.10) ; 2 (0.30) ; 3 (0.30); 4 (0.20); 5 (0.10)} |
| C | [200,400] | {[0, 10[, 0.75; [10, 30[, 0.20; [30, 60], 0.05} | {1 (0.05) ; 2 (0.10) ; 3 (0.15); 4 (0.40); 5 (0.30)} |

**TABLE 9** | Quartile Representation of Airport Data in Table 8.

| Airport | Number Passengers | Delay Time on Arrival (min) | Distance Category |
|---|---|---|---|
| A | (150, 162.5, 175, 187.5, 200) | (0, 10, 17.7, 25.4, 60) | (1, 1, 2, 2, 3) |
| B | (180, 210, 240, 270, 300) | (0, 10.6, 13.4, 30, 60) | (1, 2, 3, 4, 5) |
| C | (200, 250, 300, 350, 400) | (0, 3.3, 6.7, 10, 60) | (1, 3, 4, 5, 5) |

interval-valued variables, a distribution is assumed within each observed interval, e.g., uniform or other; for a histogram-valued variable, quantiles of any histogram may be obtained by simply interpolation, assuming uniformity within each class; for categorical nominal multi-valued variables, quantiles are determined from a ranking defined on the categories based, e.g., on their frequencies, whereas in the ordinal case the ranking is given *a priori*. Having a common representation setup then allows for a unified analysis of the data set by simultaneously taking into account variables of different types.

*Example:* Consider again data for arrival flights at three airports, and that for each flight, the number of passengers, the delay time (in min), and the distance category (say, from 1 domestic flight to 5 very long-distance intercontinental flight) has been recorded. Data for each flight have then been aggregated by airport leading to the symbolic data array in Table 8.

To obtain a homogeneous data array, a quartile representation may be considered for each observation; here we assume a uniform distribution within each observed interval of variable 'Number of passengers' and within each subinterval of variable 'Delay time'. Table 9 displays the obtained quantile representation for the three airports.

## Other Types of Symbolic Data

### *Taxonomic Variables*

A variable $Y \rightarrow O$ is a taxonomic variable if $O$ has a tree structure. Taxonomies may be taken into account in obtaining descriptions of aggregated data: first, values are recorded as in the case of categorical multi-valued variables and then each set of values of $O$ is replaced by the lowest value $h$ in the taxonomy covering the values of the given set. Generally, when at least two successors of a given level $h$ are present, they may be replaced by $h$.

### *Constrained Variables*

A variable $Y'$ is hierarchically dependent from a variable $Y$ if its application is constrained by the values taken by $Y$: $Y'$ cannot be applied if $Y$ takes values within a given set $C$. In other words, a variable $Y'$ is hierarchically dependent on a variable $Y$ if $Y'$ makes no sense for some of the values $Y$ may take, and hence becomes 'nonapplicable'. For instance, if a survey contains an item on the unemployment time of a person, the variable does not apply for a person who has never been unemployed. Descriptions that do not comply with a rule are called 'noncoherent'. The consideration of hierarchical rules in SDA has been widely studied in Refs 16–18.

## FROM CLASSICAL TO SYMBOLIC DATA: HOW LARGE IS THE STEP?

SDA developed from the need to consider data that go beyond the classical model, where each 'individual' takes exactly one value per variable. To represent data taking into account the variability intrinsic to each observation, variables have been defined whose values assume new forms. The question then arises whether we are in the same data analysis framework when we allow for the variables to take multiple values. As expected, definitions of basic statistical notions do not apply automatically, and well-established properties are no longer straightforward. To apply statistical and multivariate data analysis techniques to symbolic data requires proper consideration and often the design of appropriate tools.

Consider the case of numerical variables, where the evaluation of dispersion is a central question,

  

and the consequences of different possible choices in the design of multivariate methods have to be addressed. Dispersion is important, e.g., in clustering, as the result of any clustering method depends heavily on the scales used for the variables; therefore usually data standardization must be performed prior to a clustering method. Also, many multivariate methodologies are defined by linear combinations of the descriptive variables and on the properties of dispersion measures under linear transformations. The question then arises of how should a linear combination of symbolic numerical variables be defined and which properties remain valid.

Different approaches have been considered by various authors to address these and other questions and to propose symbolic extensions of statistical multivariate data analysis methods. Most existing methods for the analysis of such data still rely on nonparametric descriptive approaches. However, recently, probabilistic approaches are being studied and developed (see, e.g., Refs 6, 19, 20) opening new paths: statistical modeling of symbolic variables then allows for estimation and hypothesis testing.

## METHODS FOR THE ANALYSIS OF SYMBOLIC DATA

In recent years, different approaches have been investigated and many methods have been proposed for the analysis of symbolic data. This, however, has not happened uniformly across data types and analysis methods: interval data are by far the most considered case and for which more methods have been developed. Cluster analysis has received considerably more attention than other multivariate methodologies.

Next, we present a nonexhaustive survey on different analysis methodologies, referring to the most important—or the first proposed—methods in each case. In a new and dynamical field of research as SDA, much work is currently being developed, so that the reader is invited to search for alternative methods in any particular field of his interest.

### Evaluating Dissimilarity with Symbolic Data

Many multivariate methods rely on dissimilarities between the entities under analysis. Several 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. Those, of course, differ according to the type of variables.

When the entities under analysis are described by interval-valued variables, specific distance measures for comparing interval-valued observations may be used. The most common are Minkowski-type distances, which result from embedding intervals in $\mathbb{R}^2$, where one dimension is used for the lower and the other for the upper bound of the intervals, and the Hausdorff distance, which evaluates the maximum distance of a set to the nearest point in the other set, i.e., two sets are close in terms of the Hausdorff distance if every point of either set is close to some point of the other set. The Mahalanobis distance between two entities $s_{i_1}, s_{i_2}$ described by vectors of intervals is defined in Ref 21 on the basis of the vectors of observed lower and upper bounds.

In Ref 22, a study is presented of different measures for comparing probability distributions, e.g., Discrepancy, Hellinger distance, Relative entropy (or Kullback–Leibler divergence), Kolmogorov (or Uniform) metric, Lévy metric, Prokhorov metric, separation distance, total variation distance, Wasserstein (or Kantorovich) metric, and $\chi^2$ distance. Among these, the Wasserstein distance and its '$L_2$' counterpart, the Mallows distance, are the most widely used for histogram-valued data.

For an overview and discussion on different alternatives, the reader may also refer to Refs 17, 23–29.

### Clustering

A wide variety of methods have been proposed for clustering symbolic data, mostly relying on distances appropriate for the data type at hand. These include partitioning *k*-means based approaches—see, e.g., Refs 18, 21, 30–35—and the corresponding fuzzy extensions—Refs 36–40, and/or using adaptive distances—Refs 41–44. In Ref 45, the authors use the Mahalanobis–Wasserstein distance to define a new *k*-means-type method for histogram-valued data. Hierarchical clustering based on classical aggregation indices has been addressed in Ref 46 and, in Ref 47, the authors propose an extension of the Ward method. Irpino and Verde[48] successfully used the Mahalanobis–Wasserstein distance for hierarchical clustering. Brito[49] proposed a conceptual clustering approach, using the hierarchical and pyramidal models; this approach has been extended to distributional variables in Ref 12 and to constrained symbolic data in Ref 50; more recently, Brito and Polaillon[51] further developed the method using interval-based generalisation. In Ref 52, a divisive hierarchical clustering method is proposed for interval and categorical modal variables that produces 'monothetic' clusters,

  

i.e., each cluster formed is associated with a conjunction of properties on the descriptive variables, constituting a necessary and sufficient condition for cluster membership. Brito and Chavent[54] extend the divisive algorithm proposed in Refs 52, 53 to data described by interval and/or histogram-valued variables. Brito and Ichino,[55,56] are developing hierarchical clustering methods based on quantile representations of the data.[14,15] Having a common representation setup based on a quantile-vector representation (for a pre-chosen set of quantiles) allows for a unified analysis of the data set by taking simultaneously into account variables of different types. Self-Organizing Maps methodologies have been developed by Bock,[57,58]; in Ref 59 the authors introduce a batch self-organizing map algorithm based on adaptive distances; in Ref 60 an adaptive batch SOM method for Multiple Dissimilarity Data Tables is proposed; other approaches are investigated by Hajjar and Hamdan (see e.g., Refs 61, 62) and Yang et al.[63] Recently, a model-based approach for clustering interval data has been developed,[64] extending the Gaussian models proposed in Ref 6 to the model-based clustering context. For this purpose, the EM algorithm has been adapted for different covariance configurations.

## Classification

In Ref 65, different approaches to discriminant analysis of interval data are compared and classification is performed using distance-based methods. More recently, these authors have developed parametric discriminant rules, based on the models proposed in Ref 6. A thorough comparison based on an extensive simulation study is carried out in Ref 66, concluding that, in general, parametric methods outperform distance-based ones.

A method—named 'TREE'—for building a decision tree on distributional data is developed in Ref 67; it may also apply to other variable types (interval or multi-valued) variables by transforming them to distributional variables with uniform probability functions (for this methodology, see also Ref 68).

In Ref 27, the authors study the performance of the $k$ nearest neighbor method using different distance measures and for different types of symbolic data; data sets with set-valued variables (of different types), distributional-valued variables, or both, are analyzed. The proposed method is quite adaptive; it has moreover the advantage of allowing estimating the target class locally and differently for each element to be classified. Finally, the method provides a distribution of class labels for each element rather than a single value.

For the particular case of interval-valued data, different classification approaches have been investigated by several authors. Ishibuchi, Tanaka and Noriko Fukuoka[69] determine interval representations in a discriminant space using a mathematical programming formulation; the proposed method is applied to a chemical sensing problem. Jahanshahlooa et al.[70] develop a data envelopment analysis–discriminant analysis methodology for interval data, using mathematical programming and goal programming. Approaches of discriminant analysis of interval data based on imprecise probability theory may be found in Refs 71, 72. In Ref 73, a generalization of classical Factorial Discriminant Analysis to symbolic data is proposed. The method is based on a numerical analysis of the transformed symbolic data array, followed by a symbolic interpretation of the results; it allows considering numerical, categorical nominal, or distributional variables; classification rules are based on proximities in the factorial plane (see also Ref 74). Bayesian decision trees for the case when predictors are interval variables are presented in Ref 75. Discriminant analysis of interval data has also been addressed using Support Vector Machines—see Refs 76–78—and Artificial Neural Networks, in Refs 79–82.

## Factorial Analysis

Principal Component Analysis (PCA) of interval data has first been addressed in Refs 83 and 84. representing the observed intervals by their centers—'centers method'—or by considering all the vertices of the hypercube representing each of the $n$ entities in a $p$-dimensional space—the 'vertices method'. In Ref 85, a different approach is followed, where each variable is represented by the midpoints and ranges of its interval values. Three methods for Principal Component Analysis of fuzzy interval data are discussed in Ref 86. Zuccolotto[87] uses a symbolic data approach for PCA of data described by the estimated means of a $p$-dimensional variable. Rodriguez and collaborators,[88,89] have extended PCA to histogram-valued variables by representing each observation by a succession of $k$ interval nested entities ($k$ being the maximum number of categories). Ichino[15] has developed a method for Symbolic PCA of histogram-valued data based on a quantile representation where the degree of similarity between variables is evaluated by the Kendall or the Spearman's rank correlation coefficient, and then a traditional PCA is executed. Makosso-Kallyth and Diday[90] proposed an adaptation of interval PCA to symbolic histogram variables. In Ref 91, Generalized Canonical Analysis

  

is investigated, which allows for the factorial analysis of different variable types (interval, categorical multi-valued, modal).

## Multiple Regression

Linear Regression in SDA has firstly been addressed for the case of interval variables. Billard and Diday[92] proposed the first linear regression model for interval-valued variables, using the empirical covariance obtained assuming uniformity in each observed interval; 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. Neto and De Carvalho[93] have proposed a different 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; later in Ref 94, the same authors proposed a new model with non-negativity constraints on the midranges regression coefficients.

As concerns histogram-valued variables, the model in Ref 95, an extension of the authors' first model, uses the obtained covariance values for the estimation of the regression coefficients.

Irpino and Verde[96] have developed a simple linear regression model for histogram-valued data, which minimizes the Mallows' distance between the observed quantile function of the dependent variable and the one derived from the linear model. The proposed method lies in the exploitation of the properties of a decomposition of the Wasserstein distance obtained by Irpino and Romano[97]; this is used to measure the sum of squared errors and rewrite the model splitting the contribution of the predictors in a part depending on the averages of the distributions and another depending on the centered quantile distributions.

Dias and Brito[98] proposed the *Distribution and Symmetric Distribution* Linear Regression model. The distributions taken by the histogram-valued variables are represented by their quantile function, as this is the representation used by the error measure, i.e., the Mallows distance. The model includes both the quantile functions that represent the distributions that the independent histogram-valued variables take, and the quantile functions that represent the distributions that the respective symmetric histogram-valued variables take. Therefore, although non-negativity constraints on the parameters are imposed, this does not imply a direct linear relationship. Determination of the model requires solving a quadratic optimization problem, subject to non-negativity constraints on the unknowns.

## Time Series Analysis

The first work to consider interval-valued time series data used an approach based on fitting univariate ARIMA processes to the interval bounds.[99] Maia et al.[100] proposed fitting univariate ARIMA processes to the midpoint and radius and using them to forecast the interval bounds, as well as an approach based on an artificial neural network model and a combination of the two. In Refs 101–105, the authors define interval stochastic processes, interval-valued time series, weak stationarity for interval processes and, based on the sample moments previously proposed, define the empirical autocovariance and autocorrelation functions for interval time series data. In Refs 101–103, forecasting is mainly based on Vector Autoregressive models (VAR models), Vector Error Correction models (VEC), and smoothing filters; in Ref 102, VAR and interval Multilayer Perceptrons are compared.

More recently, Teles and Brito[106] proposed modeling interval-time series with Space-Time Autoregressive models.

Forecasting when data are described by histogram-valued variables has been addressed in Ref 101, see also Ref 107; in Refs 103, 104 financial applications are presented. The Mallows or Wasserstein distances are used to obtain the mean error between observed and forecast histogram values.

## Formal Concept Analysis

The importance of Galois lattices for data analysis has been widely recognized, and their use in Data Mining applications is nowadays quite frequent. The definition and use of Galois connections and Galois lattices for symbolic data has been addressed in Ref 49, developing appropriate operators for the new variable types; this has then been further developed in Refs 108, 109. Brito and Polaillon[110] define two Galois connections on a set of distributional data and the corresponding concept lattices, which use, respectively, the Maximum and the Minimum operators for generalization. Recently, the same authors[111] have proposed a novel approach, which determines intents by intervals, thereby producing more homogeneous concepts, which are easier to interpret.

# CONCLUSION

SDA extends classical Statistics, Multivariate Data Analysis and Data Mining to more complex data, in a framework which allows taking into account variability inherent to the data. The newly defined variables assume new types of realizations, which appropriately

  

model these type of data, thereby avoiding unnecessary loss of information. In recent years, different approaches have been investigated and many methods proposed for the analysis of symbolic data. Applications in different domains have shown the usefulness of the proposed approaches. Much remains however to be done. We have just rather briefly mentioned some of the issues that arise when we leave the classical data framework and allow for more complex variable types. Usual properties may not be taken for granted, and often new concepts must be designed. Schweizer wrote, back in 1984, that 'Distributions are the numbers of the future'—and perhaps this summarizes the approach presented here. For sure, a broad world of problems remains open, waiting to be explored.

SOFTWARE

To aggregate 'microdata' in a convenient way and analyze symbolic data, different software packages exist today.

- SODAS—a free package, although registration required and a code needed for installation, see: http://www.info.fundp.ac.be/asso/sodaslink.htm
- SYR—a commercial software, see: http://syrokko.com/
- R packages are available on CRAN, these are increasing, but so far mainly (though not only) for interval data. A nonexhaustive list includes: RSDA (R to SDA), MAINT.Data (Model and Analyze Interval Data), ISDA.R (interval symbolic data analysis for R), iRegression (Regression methods for interval-valued variables).

## ACKNOWLEDGMENTS

This research is supported by the Project NORTE-07-0124-FEDER-000059 within the North Portugal Regional Operational Programme (ON.2—O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT).

## REFERENCES

1. Noirhomme-Fraiture M, Brito P. Far beyond the classical data models: symbolic data analysis. *Stat Anal Data Min* 2011, 4:157–170.
2. Pedrycz W. *Granular Computing: Analysis and Design of Intelligent Systems*. Boca Raton, FL: CRC Press/Taylor & Francis; 2013.
3. Balzanella A, Rivoli L, Verde R. Data stream summarization by histograms clustering. In: Giudici P, Ingrassia S, Vichi M, eds. *Statistical Models for Data Analysis*. Berlin/Heidelberg: Springer; 2013, 27–35.
4. Giordano G, Brito P. Social networks as symbolic data. In: Vicari D, Okada A, Ragozini G, Weihs C, eds. *Analysis and Modeling of Complex Data in Behavioral and Social Sciences*. Berlin/Heidelberg: Springer; 2014.
5. Bock H-H. Symbolic data. In: Bock H-H, Diday E, eds. *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Berlin/Heidelberg: Springer; 2000, 39–53.
6. Brito P, Duarte Silva AP. Modelling interval data with normal and skew-normal distributions. *J Appl Stat* 2012, 39:3–20.
7. Duarte Silva AP, Brito P. MAINT.DATA: model and analyze interval data. R Package, version 0.2. Available at: http://cran.r-project.org/web/packages/MAINT.Data/index.html. (Accessed August 2014)
8. Choquet G. Theory of capacities. *Ann Institut Fourier* 1954, 5:131–295.
9. Dubois D, Prade H. Properties of measures of information in evidence and possibility theories. *Fuzzy Sets Syst* 1999, 100:35–49.
10. Walley P. Towards a unified theory of imprecise probability. *Int J Approx Reason* 2000, 24:125–148.
11. Diday E. Probabilist, possibilist and belief objects for knowledge analysis. *Ann Oper Res* 1995, 55:227–276.
12. Brito P. Symbolic clustering of probabilistic data. In: Rizzi A, Vichi M, Bock H-H, eds. *Advances in Data Science and Classification*. Berlin/Heidelberg: Springer; 1998, 385–389.
13. Brito P, De Carvalho FAT. Hierarchical and pyramidal clustering. In: Diday E, Noirhomme-Fraiture M, eds. *Symbolic Data Analysis and the Sodas Software*. Chichester, UK: John Wiley & Sons; 2008, 181–203.
14. Ichino M. Symbolic PCA for histogram-valued data. In: *Proceedings of the IASC 2008*, Yokohama, Japan, 2008.
15. Ichino M. The quantile method for symbolic principal component analysis. *Stat Anal Data Min* 2011, 4:184–198.
16. Vignes R. *Caractérisation Automatique de Groupes Biologiques*. PhD Thesis, University Paris VI, 1991.

  

17. De Carvalho FAT. Proximity coefficients between boolean symbolic objects. In: Diday E, Lechevallier Y, Schader M, Bertrand P, Burtschy B, eds. *New Approaches in Classification and Data Analysis*. Berlin/Heidelberg: Springer; 1994, 387–394.
18. Csernel M, De Carvalho FAT. Usual operations with symbolic data under Normal Symbolic Form. *Appl Stoch Model Bus Ind* 1999, 15:241–257.
19. Neto EAL, Cordeiro GM, de Carvalho FAT. Bivariate symbolic regression models for interval-valued variables. *J Stat Comput Simul* 2011, 81:1727–1744.
20. Le-Rademacher J, Billard L. Likelihood functions and some maximum likelihood estimators for symbolic data. *J Stat Plan Inference* 2011, 141:1593–1602.
21. De Souza RMCR, De Carvalho FAT, Tenorio CP. Two partitional methods for interval-valued data using Mahalanobis distances. In: Lemaître C, García CAR, Jesús AG, eds. *Advances in Artificial Intelligence-Proceedings IBERAMIA 2004*, 9th Ibero-American Conference on AI, Puebla, México: Springer Lecture Notes in Computer Science; 2004, 454–463.
22. Gibbs AL, Su FE. On choosing and bounding probability metrics. *Int Stat Rev* 2002, 70:419–435.
23. Bock H-H. Dissimilarity measures for probability distributions. In: Bock H-H, Diday E, eds. *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Berlin/Heidelberg: Springer; 2000, 153–165.
24. Malerba D, Esposito F, Gioviale V, Tamma V. Comparing dissimilarity measures in symbolic data analysis. In: *Pre-Proceedings of EKT-NTTS*, Vol. 1, 2001, 473–481.
25. Malerba D, Esposito F, Monopoli M. Comparing dissimilarity measures for probabilistic symbolic objects. In: Zanasi A, Brebbia CA, Ebecken NFF, Melli P, eds. *Data Mining III, Series Management Information Systems*, vol. 6. Southampton, UK: WIT Press; 2002, 31–40.
26. Diday E, Esposito F. An introduction to symbolic data analysis and the SODAS software. *Intell Data Anal* 2003, 7:583–602.
27. Appice A, D'Amato C, Esposito F, Malerba D. Classification of symbolic objects: a lazy learning approach. *Intell Data Anal* 2006, 10:301–324.
28. Verde R, Irpino A. Dynamic clustering of histogram data: using the right metric. In: Brito P, Bertrand P, Cucumel G, De Carvalho F, eds. *Selected Contributions in Data Analysis and Classification*. Berlin/Heidelberg: Springer; 2007, 123–134.
29. Esposito F, Malerba D, Appice A. Dissimilarity and matching. In: Diday E, Noirhomme-Fraiture M, eds. *Symbolic Data Analysis and the Sodas Software*. Chichester, UK: John Wiley & Sons; 2008, 123–148.
30. De Souza RMCR, De Carvalho FAT. Dynamic clustering of interval data based on adaptive Chebyshev distances. *Electron Lett* 2004, 40:658–659.
31. De Souza RMCR, De Carvalho FAT. Clustering of interval data based on city-block distances. *Pattern Recogn Lett* 2004, 25:353–365.
32. De Carvalho FAT, Tenorio C, Lechevallier Y. Dynamic cluster methods for interval data based on Mahalanobis distances. In: Banks D, McMorris FR, Arabie P, Gaul W, eds. *Classification, Clustering, and Data Mining Applications – Proceedings of the IFCS 2004*. Berlin/Heidelberg: Springer; 2004, 351–360.
33. De Carvalho FAT, Brito P, Bock H-H. Dynamic clustering for interval data based on $L_2$ distance. *Comput Stat* 2006, 21:231–250.
34. Chavent M, Lechevallier Y, Verde R. New clustering methods for interval data. *Comput Stat* 2006, 21:211–229.
35. De Carvalho FAT, De Souza RMCR. Unsupervised pattern recognition models for mixed feature-type symbolic data. *Pattern Recogn Lett* 2010, 31:430–443.
36. El-Sonbaty Y, Ismail MA. Fuzzy clustering for symbolic data. *IEEE Trans Fuzzy Syst* 1998, 6:195–204.
37. Yang M-S, Hwang P-Y, Chen D-H. Fuzzy clustering algorithms for mixed feature variables. *Fuzzy Sets Syst* 2004, 141:301–317.
38. D'Urso P, Giordani P. A weighted fuzzy c-means clustering model for fuzzy data. *Comput Stat Data Anal* 2006, 50:1496–1523.
39. De Carvalho FAT. Fuzzy c-means clustering methods for symbolic interval data. *Pattern Recogn Lett* 2007, 28:423–437.
40. Jeng J-T, Chuan C-C, Tseng C-C, Juan C-J. Robust interval competitive agglomeration clustering algorithm with outliers. *Int J Fuzzy Syst* 2010, 12: 227–236.
41. De Souza RMCR, De Carvalho FAT, Silva FCD. Clustering of interval-valued data using adaptive squared Euclidean distance. In: Pal NR, Kasabov N, Mudi RK, Pa S, Paruil SK, eds. *Neural Information Processing – Proceedings of the ICONIP*. Berlin/Heidelberg: Springer; 2004, 775–780.
42. De Carvalho FAT, De Souza RMCR, Chavent M, Lechevallier Y. Adaptive Hausdorff distances and dynamic clustering of symbolic interval data. *Pattern Recogn Lett* 2006, 27:167–179.
43. De Carvalho FAT, Lechevallier Y. Partitional clustering algorithms for symbolic interval data based on single adaptive distances. *Pattern Recogn* 2009, 42:1223–1236.
44. De Carvalho FAT, Tenorio CP. Fuzzy k-means clustering algorithms for interval-valued data based on adaptive quadratic distances. *Fuzzy Sets Syst* 2010, 161:2978–2999.
45. Verde R, Irpino A. Comparing histogram data using a Mahalanobis-Wasserstein distance. In: Brito P,


  

ed. *Proceedings of the COMPSTAT'2008*. Berlin/Heidelberg: Springer; 2008, 77–89.

46. Hardy A, Lallemand P. Clustering of symbolic objects described by multi-valued and modal variables. In: Banks D, McMorris F, Arabie P, Gaul W, eds. *Classification, Clustering, and Data Mining Applications*. Berlin/Heidelberg: Springer; 2004, 325–332.
47. Korenjak-Cerne S, Batagelj V, Japelj Pavešic B. Clustering large data sets described with discrete distributions and its application on TIMSS data set. *Stat Anal Data Min* 2011, 4:199–215.
48. Irpino A, Verde R. A New Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In: Batagelj V, Bock H-H, Ferligoj A, Žiberna A, eds. *Proceedings of the IFCS 2006*. Berlin/Heidelberg: Springer; 2006, 185–192.
49. Brito P. Symbolic objects: order structure and pyramidal clustering. *Ann Oper Res* 1995, 55:277–297.
50. Brito P, De Carvalho FAT. Symbolic clustering of constrained probabilistic data. In: Opitz O, Schwaiger M, eds. *Exploratory Data Analysis in Empirical Research*. Berlin/Heidelberg: Springer; 2002, 12–21.
51. Brito P, Polaillon G. Classification conceptuelle avec généralisation par intervalles [Conceptual clustering with generalization by intervals]. *Revue des Nouvelles Technologies de l'Information* 2012, E.23:35–40.
52. Chavent M. A monothetic clustering method. *Pattern Recogn Lett* 1998, 19:989–996.
53. Chavent M, Lechevallier Y, Briant O. DIVCLUS-T: a monothetic divisive hierarchical clustering method. *Comput Stat Data Anal* 2007, 52:687–701.
54. Brito P, Chavent M. Divisive monothetic clustering for interval and histogram- valued data. In: *Proceedings of the ICPRAM 2012 – 1st International Conference on Pattern Recognition Applications and Methods*, Vilamoura, Portugal, 2012.
55. Brito P, Ichino M. Symbolic clustering based on quantile representation. In: Lechevallier Y, Saporta G, eds. *Proceedings of the COMPSTAT 2010*. Heidelberg: Physica Verlag; 2010.
56. Brito P, Ichino M. Clustering symbolic data based on quantile representation. In: Brito P, Noirhomme-Fraiture M, eds. *Proceedings of the Workshop in Symbolic Data Analysis*. Belgium: Namur; 2011.
57. Bock H-H. Clustering methods and Kohonen maps for symbolic data. *J Jpn Soc Comput Stat* 2002, 15:217–229.
58. Bock H-H. Visualizing symbolic data by Kohonen maps. In: Diday E, Noirhomme-Fraiture M, eds. *Symbolic Data Analysis and the SODAS Software*. Chichester, UK: John Wiley & Sons; 2008, 205–234.
59. Pacifico LDS, De Carvalho FAT. A batch self-organizing maps algorithm based on adaptive distances. In: *Proceedings of the 2011 International Joint Conference on Neural Networks (IJCNN)*, San Jose, CA: IEEE; 2011, 2297–2304.
60. Dos S, Dantas AB, De Carvalho FAT. Adaptive batch SOM for multiple dissimilarity data tables. In: *Proceedings of the 23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, Boca Raton, FL: IEEE; 2011, 575–578.
61. Hajjar C, Hamdan H. Self-organizing map based on Hausdorff distance for interval-valued data. In: *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Anchorage, AK: IEEE; 2011a, 1747–1752.
62. Hajjar C, Hamdan H. Self-organizing map based on $L_2$ distance for interval-valued data. In: *Proceedings of the 6th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI)*, Timisoara: IEEE; 2011b, 317–322.
63. Yang M-S, Hung W-L, Chen D-H. Self-organizing map for symbolic data. *Fuzzy Sets Syst* 2012, 203:49–73.
64. Brito P, Duarte Silva AP, Dias JG. Probabilistic clustering of interval data. *Intell Data Anal*. In Press.
65. Duarte Silva AP, Brito P. Linear discriminant analysis for interval data. *Comput Stat* 2006, 21:289–308.
66. Duarte Silva AP, Brito P. Discriminant analysis of interval data: parametric versus distance-based approaches. Under Revision.
67. Périnel E, Lechevallier Y. Symbolic discrimination rules. In: Bock H-H, Diday E, eds. *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*. Berlin/Heidelberg: Springer; 2000, 244–265.
68. Ciampi A, Diday E, Lebbe J, Périnel E, Vignes R. Growing a tree classifier with imprecise data. *Pattern Recogn Lett* 2000, 21:787–803.
69. Ishibuchi H, Tanaka H, Noriko Fukuoka N. Discriminant analysis of multi-dimensional interval data and its application to chemical sensing. *Int J Gen Syst* 1990, 16:311–329.
70. Jahanshahlooa GR, Lotfib FH, Balfc FR, Rezaid HZ. Discriminant analysis of interval data using Monte Carlo method in assessment of overlap. *Appl Math Comput* 2007, 191:521–532.
71. Nivlet P, Fournier F, Royer JJ. Interval discriminant analysis: an efficient method to integrate errors in supervised pattern recognition. In: *Proceedings ISIPTA 2001*, Second international symposium on imprecise probabilities and their applications, Cornell University, Ithaca, NY, USA, 2001, 284–292.
72. Utkin LV, Coolen FPA. Interval-valued regression and classification models in the framework of machine learning. In: *Proceedings of the 7th International Symposium on Imprecise Probability: Theories and Applications*, Innsbruck, Austria, 2011.
73. Lauro NC, Verde R, Palumbo F. Factorial discriminant analysis on symbolic objects. In: Bock H-H, Diday E,

  

eds. *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*. Berlin/Heidelberg: Springer; 2000, 212–233.

74. Lauro NC, Verde R, Irpino A. Factorial discriminant analysis. In: Diday E, Noirhomme-Fraiture M, eds. *Symbolic Data Analysis and the Sodas Software*. Chichester, UK: John Wiley & Sons; 2008, 341–358.
75. Rasson JP, Pirçon J-Y, Lallemand P, Adans S. Unsupervised divisive classification. In: Diday E, Noirhomme-Fraiture M, eds. *Symbolic Data Analysis and the Sodas Software*. Chichester, UK: John Wiley & Sons; 2008, 149–156.
76. Do T-N, Poulet F. Kernel methods and visualization for interval data mining. In: Janssen J, Lenca P, eds. *Proceedings of the Conference on Applied Stochastic Models and Data Analysis, ASMDA 2005*. Bretagne, France: ENST; 2005.
77. Carrizosa E, Gordillo J, Plastria F. Classification problems with imprecise data through separating hyperplanes. Available at: http://www.optimization-online.org/DBFILE/2007/09/1781.pdf. (Accessed August 2014)
78. Angulo C, Anguita D, González L. Interval discriminant analysis using support vector machines. In: *ESANN'2007 Proceedings – European Symposium on Artificial Neural Networks*, Bruges, Belgium, 2007.
79. Síma J. Neural expert systems. *Neural Netw* 1995, 8:261–271.
80. Simoff SJ. Handling uncertainty in neural networks: an interval approach. In: *Proceedings of the IEEE International Conference on Neural Networks*. Washington DC: IEEE Computer Society Press; 1996, 606–610.
81. Beheshti M, Berrached A, de Korvin A, Hu C, Sirisaengtaksin O. On interval weighted freelayer neural networks. In: *Proceedings of the 31st Annual Simulation Symposium*. Washington DC: IEEE Computer Society Press; 1998, 188–194.
82. Rossi F, Conan Guez B. Multilayer perceptron on interval data. In: Jajuga K, Sokolowski A, Bock H-H, eds. *Classification, Clustering and Data Analysis*. Berlin/Heidelberg: Springer; 2002, 427–434.
83. Chouakria A, Cazes P, Diday E. Symbolic Principal Component Analysis. In: Bock H-H, Diday E, eds. *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*. Berlin/Heidelberg: Springer; 2000, 200–212.
84. Cazes P, Chouakria A, Diday E, Schektman Y. Extensions de l'Analyse en Composantes Principales à des données de type intervalle. *Revue de Statistique Appliquée* 1997, 24:5–24.
85. Lauro C, Palumbo P. Principal Component Analysis for non-precise data. In: Vichi M, Monari P, Mignani S, Montanari A, eds. *New Developments in Classification and Data Analysis*. Berlin/Heidelberg: Springer; 2005, 173–184.
86. Giordani P, Kiers HAL. A comparison of three methods for Principal Component Analysis of fuzzy interval data. *Comput Stat Data Anal, Special Issue Fuzzy Approach to Statistical Analysis* 2006, 51:379–397 .
87. Zuccolotto P. Principal components of sample estimates: an approach through symbolic data analysis. *Stat Meth Appli* 2007, 16:173–192.
88. Rodriguez O, Diday E, Winsberg S. Generalization of the Principal Component Analysis to histogram data. In: *Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Data Bases, Workshop on Symbolic Data Analysis*, Lyon, France, 2000.
89. Rodriguez O, Pacheco A. Applications of histogram Principal Component Analysis. In: *Proceedings of the 15th European Conference on Machine Learning (ECML) and the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, Pisa, Italy, 2004.
90. Makosso-Kallyth S, Diday E. Adaptation of interval PCA to symbolic histogram variables. *Adv Data Anal Classif* 2012, 6:147–159.
91. Lauro C, Verde R, Irpino A. Generalized canonical analysis. In: Diday E, Noirhomme-Fraiture M, eds. *Symbolic Data Analysis and the Sodas Software*. Chichester, UK: John Wiley & Sons; 2008, 313–330.
92. Billard L, Diday E. Regression analysis for interval-valued data. In: *Data Analysis, Classification, and Related Methods, Proceedings of the Seventh Conference of the International Federation of Classification Societies (IFCS00)*. Berlin/Heidelberg: Springer; 2000, 369–374.
93. Neto EAL, De Carvalho FAT. Centre and range method for fitting a linear regression model to symbolic interval data. *Comput Stat Data Anal* 2008, 52:1500–1515.
94. Neto EAL, De Carvalho FAT. Constrained linear regression models for symbolic interval-valued variables. *Comput Stat Data Anal* 2010, 54:333–347.
95. Billard L, Diday E. Symbolic regression analysis. In: *Classification, Clustering and Data Analysis, Proceedings of the Conference of the International Federation of Classification Societies (IFCS02)*. Berlin/Heidelberg: Springer; 2002, 281–288.
96. Verde R, Irpino A. Ordinary least squares for histogram data based on Wasserstein distance. In: Lechevallier Y, Saporta G, eds. *Proceedings of the COMPSTAT'2010*. Heidelberg: Physica Verlag; 2010, 581–589.
97. Irpino A, Romano E. Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation. In: Noirhomme-Fraiture M, Venturini G, eds. *EGC, RNTI-E-9 of Revue des Nouvelles Technologies de l'Information*. Toulouse, France: Cépaduès-Éditions; 2007, 99–110.

  

98. Dias S, Brito P. Distribution and symmetric distribution regression model for histogram-valued variables. Submitted. arXiv:1303.6199v1 [stat.ME].

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

100. Maia ALS, De Carvalho FAT, Ludermir TD. Forecasting models for interval-valued time series. *Neurocomputing* 2008, 71:3344–3352.

101. Arroyo J. *Métodos de Predicción para Series Temporales de Intervalos e Histogramas*, PhD Thesis, Universidad Pontifícia Comillas, Madrid, Spain, 2008.

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

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

104. González-Rivera G, Arroyo J. Time series modeling of histogram-valued data: the daily histogram time series of S&P500 intradaily returns. *Int J Forecast* 2012, 28:20–33.

105. Han A, Hong Y, Lai K, Wang S. Interval time series analysis with an application to the Sterling-Dollar exchange rate. *J Syst Sci Complex* 2008, 21:558–573.

106. Teles P, Brito P. Modelling interval time series with Space-time processes. *Commun Stat Theor Meth* (In press). doi: 10.1080/03610926.2013.782200.

107. Arroyo J, Maté C. Forecasting histogram time series with k-nearest neighbours methods. *Int J Forecast* 2009, 25:182–207.

108. Polaillon G. Interpretation and reduction of Galois lattices of complex data. In: Rizzi A, Vichi M, Bock H-H, eds. *Advances in Data Science and Classification*. Berlin/Heidelberg: Springer; 1998, 433–440.

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

110. Brito P, Polaillon G. Structuring probabilistic data by Galois lattices. *Math Social Sci* 2005, 169: 77–104.

111. Brito P, Polaillon G. Homogeneity and stability in conceptual analysis. In: Napoli A, Vychodil V, eds. *Proceedings of the 8th International Conference on Concept Lattices and Their Applications*. Nancy: INRIA; 2011, 251–263.

## FURTHER READING

Billard L, Diday E. From the statistics of data to the statistics of knowledge: Symbolic Data Analysis. *J Am Stat Assoc* 2003, 98:470–487.

Billard L, Diday E. *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Chichester, UK: John Wiley & Sons; 2006.

Bock H-H, Diday E, eds. *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*. Berlin/Heidelberg: Springer; 2000.

Brito P. On the analysis of symbolic data. In: Brito P, Bertrand P, Cucumel G, De Carvalho FAT, eds. *Selected Contributions in Data Analysis and Classification*. Berlin/Heidelberg: Springer; 2007, 13–22.

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

Diday E. Introduction à l'analyse des données symboliques. *Revue Recherche Opérationnelle* 1989, 23:193–236.

Diday E, Noirhomme-Fraiture M, eds. *Symbolic Data Analysis and the Sodas Software*. Chichester, UK: John Wiley & Sons; 2008.

  