

New skills in Symbolic Data Analysis for Official Statistics

Rosanna Verde^{*1}, Vladimir Batagelj^{2,3}, Paula Brito⁴, A. Pedro Duarte Silva⁵,
Simona Korenjak-Černe^{6,2}, Jasminka Dobša⁷, and Edwin Diday^{†8}

^{*1}DMF - University of the Campania, Italy. Corresponding author. email:
rosanna.verde@unicampania.it

²IMFM Ljubljana, Slovenia.

³University of Primorska, Koper, Slovenia.

⁴Fac. Economia, Universidade do Porto & LIAAD-INESC TEC, Portugal.

⁵Universidade Católica Portuguesa, Católica Porto Business School, Portugal.

⁶University of Ljubljana, SEB, Slovenia.

⁷University of Zagreb Faculty of Organization and Informatics, Croatia.

⁸CEREMADE Laboratory, University Paris Dauphine, Paris, France.

June 25, 2024

Abstract

The paper draws attention to the use of Symbolic Data Analysis (SDA) in the field of Official Statistics. It is composed of three sections presenting three pilot techniques in the field of SDA. The three contributions range from a technique based on the notion of exactly unified summaries for the creation of symbolic objects, a model-based approach for interval data as an innovative parametric strategy in this context, and measures of similarity defined between a class and a collection of classes based on the frequency of the categories which characterize them.

The paper shows the effectiveness of the proposed approaches as prototypes of numerous techniques developed within the SDA framework and opens to possible further developments.

Keywords— aggregate data, interval data, mergeable summaries, s-concordance, s-discordance

1 Introduction

This paper intends to refer to the contribution of Symbolic Data Analysis [1] in the domain of Official Statistics (OSs). As is well known, OSs are presented in the form of aggregate data, primarily as summary measures of complex economics and social phenomena with high variability in time and space. They are

[†]Edwin Diday deceased on the 28th April 2023. His co-authorship of this manuscript has the agreement of his heirs.

28 mostly provided as composite indicators, distributions of phenomena at the spatial level, usually referring
29 to different categories of reference populations or strata of the target population. The more recent use
30 of Big Data techniques, such as Machine Learning, increasingly requires adequate synthesis to reduce the
31 dimensionality of data. OSs are reported in the form of aggregates both as the outcome of appropriate
32 summaries and in order to preserve the confidentiality and privacy of the data. Symbolic Data Analysis was
33 initiated at the end of the 1980s with the pioneering work of Edwin Diday ([2], [3], [4]) and it has represented
34 a new branch of research with the aim of modeling statistical units no more through punctual (categorical
35 or numerical) values observed on a set of characters, and collected in a classical table ($n \times p$) of individuals x
36 variables. It has allowed defining statistical units as concepts (i.e. symbolic objects) with the possibility of
37 assuming, with respect to each variable, a plurality of values that can be observed (e.g. intervals, frequency
38 distributions, multiple categories) or defined based on *domain expert knowledge*. The concept of *categories*
39 is evoked and a statistical unit refers to a species, a family, a class, ... The description of these statistical
40 units, whether observed or conceptual (the latter, when the data come from expert knowledge or synthesis),
41 is solved into symbolic data and expressed by the realization of multiple-valued descriptors, referred to as
42 *symbolic variables*. Further, the *symbolic data table* is defined to contain a real interval, a distribution, or
43 a sequence of categories in each cell. It is also possible to consider relationships and taxonomies between
44 variables in the description of concepts. However, the relationship between the variables or the taxonomic
45 structure of the categories finds little translation in the description of the symbolic data. The symbolic
46 data defined from the multiple values for the different statistical units being analyzed thus represent a set
47 of data, in aggregate form, to be analyzed taking into account the main characteristics of the variability of
48 the description of each statistical unit and the mixed nature of the descriptors that can be both qualitative
49 and quantitative multi-valued. Moreover, the data are not in vectorial form, and this leads to the search for
50 metrics and representation spaces as well as to the construction of elementary statistics, and to the extension
51 of classical analysis techniques.

52 The development of SDA was made possible by the theoretical and conceptual framework elaborated by
53 Edwin Diday and by the involvement of numerous researchers, not only from Europe, who collaborated on
54 two major European projects between the late 1990s and early 2000s. These latter allowed for developing
55 Symbolic Data Analysis methods, standardizing data representation, and exploring applications in Official
56 Statistics. The first European research project, “Symbolic Objects Data Analysis System”, SODAS (1996-
57 1999), gathered 17 teams working in SDA, including National Statistical Institutes (NSI’s). The project led
58 to the first statistical package for SDA, ‘SODAS’, which made it possible for Data Analysis researchers and
59 users alike to produce, edit, and analyze symbolic data. At the same time, the first book on SDA “Analysis
60 of Symbolic Data” [1] was published. SODAS was followed by another European project, “Analysis System
61 of Symbolic Official Data”, ASSO, gathering 15 teams, including three NSI’s. The ASSO project allowed
62 the development of new methodologies and the publishing of a second book – “Symbolic Data Analysis and
63 the SODAS Software” [5].

64 These projects fostered the development of a field of research that, through doctoral theses, workshops,
65 conference sessions, publications in scientific journals, in the fields of Statistics, Data Analysis, and Com-
66 putational Statistics, has known considerable progress over the past twenty years, and spread far beyond
67 Europe. In the 1990s, Big Data analysis was in its early days and was not yet a challenge in the statistical
68 community, while an area that had to be dealt with by statisticians was data confidentiality and privacy.
69 It is evident that SDA has been a pioneering line of research for the analysis of Big Data, of complex, ag-
70 gregated data, which represent today, in various fields, the information to be processed. If we consider that
71 most techniques for analyzing large amounts of data are based on the use of synthesis functions, dimension

72 reduction, and analysis of aggregated data, SDA methods can provide strong support. This is demonstrated,
 73 e.g., by the use of data in the form of distributions for the synthesis of data streams or data from sensors or
 74 high-frequency time sequences. Related to the new line of research for Big Data, are the recent proposals
 75 for metrics and measures of dissimilarity to compare descriptions of aggregate data and classes. Hence,
 76 the new developments on *concordance and discordance* were introduced by the relentless research work and
 77 innovative propositions of Edwin Diday and some colleagues who have recently collaborated with him. This
 78 paper, far from summarising the enormous scope in which SDA research has developed and branched out,
 79 brings together some of the research contributions presented at the NTTS2023 conference and mentions a
 80 final scientific inheritance left by Edwin Diday.

81 The manuscript is organized as follows. In the second section, written by V. Batagelj, the notion of
 82 exactly mergeable summaries is introduced and discussed. The third section presents a contribution by P.
 83 Brito and A.P. Duarte Silva on parametric models for interval data for the discovery of patterns and trends,
 84 with application to the Portuguese Labour Force Survey. In a fourth section, provided by S. Korenjak-Černe
 85 and J. Dobša and based on the theoretical contributions of E. Diday, two examples show the applicability of
 86 concordance and discordance in two very distinct areas: one using data from the international measurement of
 87 reading achievement, and the other for the representation of textual data for automatic classification. Finally,
 88 the Conclusion section summarises the presented approaches, putting in evidence their main contributions.

89 2 Exactly mergeable summaries

90 V. Batagelj

91 2.1 Motivation

92 In our program Clamix [6] for clustering symbolic data, they are represented by discrete distributions (n, \mathbf{p})
 93 where \mathbf{p} is an empirical probability distribution and n is the number of original data units summarized by
 94 \mathbf{p} . This representation has two important properties

- 95 • fixed space required for a description of a unit/cluster;
- 96 • description of a union of two disjoint clusters can be obtained from their descriptions.

97 In this paper, we will elaborate on the second observation.

98 For example, let us consider the population pyramids of the world's countries. How to join the population
 99 pyramids of China and Vanuatu (see Fig. 1) ?

100 When comparing two countries A and B we compare the shapes of their probability distributions \mathbf{p}_A
 101 and \mathbf{p}_B . But to determine the correct probability distribution of their union $A \cup B$ we need to know also
 102 the sizes n_A and n_B of countries A and B

$$(n_{A \cup B}, \mathbf{p}_{A \cup B}) = (n_A + n_B, \frac{n_A \mathbf{p}_A + n_B \mathbf{p}_B}{n_A + n_B})$$

103 2.2 Aggregation

104 In an analysis of large data sets the *aggregation* is a standard way for reducing the size (complexity) of
 105 the data. Recently some books dealing with the theoretical and algorithmic background of the traditional
 106 aggregation (replacing values of a variable over a group by a single value) were published [7, 8, 9, 10, 11].

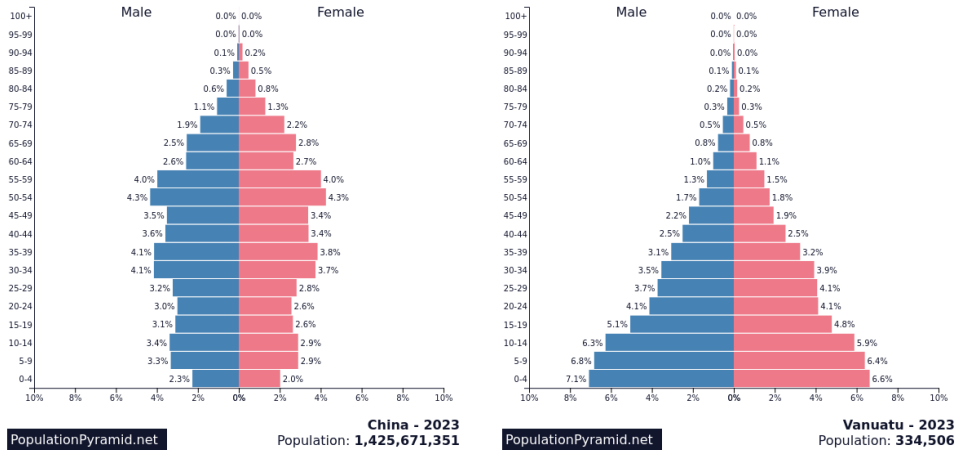


Figure 1: The population pyramids of China and Vanuatu.

107 Data analysis programs provide aggregation functions such as means (arit, geom, harm, median, modus),
 108 min, max, product, bounded sum, counting, etc. [12]. Special care has to be given to variables measured in
 109 different measurement scales.

110 In theoretical discussion the traditional aggregation functions are usually “normalized” to the interval
 111 $[0, 1]$ – they take real arguments in $[0, 1]^k$ and produce a value in $[0, 1]$, and satisfy the conditions: $f(\mathbf{0}) = 0$,
 112 $f(\mathbf{1}) = 1$, and monotonicity $\mathbf{x} \leq \mathbf{y} \Rightarrow f(\mathbf{x}) \leq f(\mathbf{y})$. Often, in applications, also idempotency and symmetry
 113 are required.

114 The applications of traditional aggregation functions are used, besides determining a representative
 115 value for a group of measurements, mainly to combine partial criteria into a single criterion (multicriteria
 116 optimization and decision-making) or to express the membership degree in combined fuzzy sets.

117 A problem with traditional aggregation is that often too much information is discarded, thus reducing
 118 the precision of the obtained results.

119 A much better, preserving more information, summarization of original data can be achieved by repre-
 120 senting aggregated data using selected types of complex data such as symbolic objects [2, 13], compositions
 121 [14], functional data [15], etc. In the Symbolic Data Analysis (SDA) framework, much work is devoted to
 122 the summarization process, for example, the function `classic.to.sym` in RSDA [16], and SODAS or SYR
 123 software.

124 2.3 Mergeable summaries

125 In complex data analysis the measured values over a selected subset of units A are aggregated into a complex
 126 object $\Sigma(A)$ and not into a single value. Most of the aggregation theory does not apply directly. In our
 127 contribution, we present an attempt to start building a theoretical background of complex aggregation.

128 An interesting question is, which complex data types are compatible with the merging of disjoint sets of
 129 units

$$\Sigma(A \cup B) = F(\Sigma(A), \Sigma(B)), \quad \text{for } A \cap B = \emptyset. \quad (1)$$

130 Selecting a name for this kind of summary we were inclined towards the term *hierarchical* or *mergeable*
 131 *summary*. Searching on Google we learned that the term *mergeable summary* was already proposed and

132 elaborated by [17]. They enable parallelization in big data algorithms and stream processing. The summa-
 133 rization in big data is not deterministic and allows some errors. A summary is *mergeable* if the error and
 134 space (size of the summary) do not increase after the merge.

135 In this paper, we will discuss exactly mergeable summaries “without errors”.

136 2.4 Exactly mergeable summaries

137 A summary $\Sigma(A)$ is an *exactly mergeable summary* if and only if it requires a fixed space of small size and
 138 satisfies the relation (1).

139 We can consider merging as a partially defined binary operation $\Sigma(A) * \Sigma(B) = F(\Sigma(A), \Sigma(B))$. For
 140 mutually disjoint subsets A, B , and C we have

$$\Sigma(A) * \Sigma(B) = \Sigma(B) * \Sigma(A)$$

141

$$\Sigma(A) * (\Sigma(B) * \Sigma(C)) = (\Sigma(A) * \Sigma(B)) * \Sigma(C)$$

142 2.4.1 Simple examples

143 We assume that a numerical variable $v : U \rightarrow \mathbb{R}$ is measured on the set of units U and that $A, B \subseteq U$ and
 144 $A \cap B = \emptyset$.

145 Let $\text{sort}_A(v)$ be a list of values of the variable v on the set of units A ordered in decreasing order. We
 146 define $\text{1st}_A(v) = \text{sort}_A(v)[1]$ and $\text{2nd}_A(v) = \text{sort}_A(v)[2]$.

147 It is easy to check that the following summaries are exactly mergeable:

- 148 1. $\Sigma(A) = |A| = n_A$
 149 $\Sigma(A \cup B) = \Sigma(A) + \Sigma(B)$
- 150 2. $\Sigma(A) = \min_{X \in A} v(X)$
 151 $\Sigma(A \cup B) = \min(\Sigma(A), \Sigma(B))$
- 152 3. $\Sigma(A) = \max_{X \in A} v(X)$
 153 $\Sigma(A \cup B) = \max(\Sigma(A), \Sigma(B))$
- 154 4. $\Sigma(A) = (\text{1st}_A(v), \text{2nd}_A(v))$
 155 $\Sigma(A \cup B) = (\text{1st}_L(v), \text{2nd}_L(v))$,
 156 where $L = \{\text{1st}_A(v), \text{2nd}_A(v), \text{1st}_B(v), \text{2nd}_B(v)\}$
 157 This example can be generalized to $\Sigma(A) = \text{Top-}k_A(v)$.
- 158 5. $\Sigma(A) = (n_A, \mu_A)$, $\mu_A = \frac{1}{n_A} \sum_{X \in A} v(X)$
 159 $\Sigma(A \cup B) = (n_A + n_B, \frac{n_A \mu_A + n_B \mu_B}{n_A + n_B})$
- 160 6. $\Sigma(A) = \sum_{X \in A} v(X)$
 161 $\Sigma(A \cup B) = \Sigma(A) + \Sigma(B)$
- 162 7. $\Sigma(A) = (n_A, \gamma_A)$, $\gamma_A = \sqrt[n_A]{\prod_{X \in A} v(X)}$
 163 $\Sigma(A \cup B) = (n_A + n_B, \sqrt[n_A + n_B]{\gamma_A^{n_A} \cdot \gamma_B^{n_B}})$

164 2.4.2 Moments

165 The distribution of values of variable v on the set of units A is often summarized by its average μ_A and its
 166 standard deviation σ_A . It would be better to represent it as $\Sigma(A) = (n_A, \mu_A, \sigma_A)$, where n_A is the number
 167 of units in A .

168 Then the distribution of additional values of variable v on the set of units B , $A \cap B = \emptyset$, is summarized
 169 by $\Sigma(B) = (n_B, \mu_B, \sigma_B)$ and can be combined into a summary of the distribution on the set $C = A \cup B$,
 170 $\Sigma(C) = (n_C, \mu_C, \sigma_C)$ determined by $\Sigma(A)$ and $\Sigma(B)$ as follows

$$n_C = n_{A \cup B} = n_A + n_B$$

$$\mu_C = \mu_{A \cup B} = \frac{n_A \mu_A + n_B \mu_B}{n_C}$$

$$\sigma_C = \sigma_{A \cup B} = \sqrt{\frac{S_C}{n_C} - \mu_C^2}$$

173 where $S_C = S_A + S_B$ and $S_X = n_X(\sigma_X^2 + \mu_X^2)$. $\Sigma(A)$ is an exactly mergeable summary.

174 This result can be extended to higher moments.

175 2.4.3 Set membership count

176 Counting the number of units from C in A

$$n(A; C) = |A \cap C|$$

177 is an exactly mergeable summary.

178 **Proof:** Since $A \cap B = \emptyset$, so is $A \cap B \cap C = \emptyset$. Therefore

$$\begin{aligned} n(A \cup B; C) &= |(A \cup B) \cap C| = |(A \cap C) \cup (B \cap C)| = \\ 179 &= |A \cap C| + |B \cap C| - |A \cap B \cap C| = n(A; C) + n(B; C) \quad \square \end{aligned}$$

180 2.4.4 Combining exactly mergeable summaries

181 Let Σ_1 and Σ_2 be exactly mergeable summaries. Then also their *composition*

$$\Sigma_1 \oplus \Sigma_2(A) = (\Sigma_1(A), \Sigma_2(A))$$

182 is an exactly mergeable summary.

183 **Proof:** $\Sigma_1 \oplus \Sigma_2(A \cup B) = (\Sigma_1(A \cup B), \Sigma_2(A \cup B)) =$

$$= (F_1(\Sigma_1(A), \Sigma_1(B)), F_2(\Sigma_2(A), \Sigma_2(B))) \quad \square$$

184 Since min and max are mergeable summaries also their composition – the *interval summary* of the
 185 variable v on the set of units A

$$\Sigma(A) = [\min_{X \in A} v(X), \max_{X \in A} v(X)]$$

186 is an exactly mergeable summary. Let $\Sigma(A) = [m_A, M_A]$ and $\Sigma(B) = [m_B, M_B]$ then

$$\Sigma(A \cup B) = [\min_{X \in A \cup B} v(X), \max_{X \in A \cup B} v(X)] = [\min(m_A, m_B), \max(M_A, M_B)] \quad \square$$

187 Let $K = \{k_1, k_2, \dots, k_s\}$ be a finite set of categories and $v : U \rightarrow K$ a categorical (nominal) variable on
 188 the set of units U . The summary

$$\Sigma(A) = \{(k, n(A; C(k))) : k \in K\} \quad \text{where} \quad C(k) = \{X : v(X) = k\}$$

189 is called a *bar chart*.

190 Let $v : U \rightarrow \mathbb{R}$ be an ordinal variable and $\mathbf{B} = (B_1, B_2, \dots, B_r)$ an ordered partition (set of bins) of
 191 $v(A)$. The summary

$$\Sigma(A) = [(B, n(A; C(B))) : B \in \mathbf{B}] \quad \text{where} \quad C(B) = \{X : v(X) \in B\}$$

192 is called a *histogram*.

193 A histogram (and also a bar chart) is essentially a frequency distribution \mathbf{f} over a given set of bins \mathbf{B}
 194 (categories K). It can be equivalently represented by a pair (n, \mathbf{p}) where $n = \sum_i f_i$ is the size of the set A
 195 and $\mathbf{p} = \frac{1}{n}\mathbf{f}$ is the corresponding probability distribution.

196 Therefore, since set membership counts are exactly mergeable, the bar charts and histograms are exactly
 197 mergeable summaries.

198 2.4.5 Proving that a summary is not exactly mergeable

199 If for a summary Σ there exist sets A_1, B_1, A_2, B_2 such that $A_1 \cap B_1 = \emptyset, A_2 \cap B_2 = \emptyset, \Sigma(A_1) = \Sigma(A_2),$
 200 $\Sigma(B_1) = \Sigma(B_2),$ and $\Sigma(A_1 \cup B_1) \neq \Sigma(A_2 \cup B_2)$ then Σ is **not** exactly mergeable.

201 **Proof:** Assume that Σ is exactly mergeable. Then

$$\Sigma(A_1 \cup B_1) = F(\Sigma(A_1), \Sigma(B_1)) = F(\Sigma(A_2), \Sigma(B_2)) = \Sigma(A_2 \cup B_2)$$

202 – a contradiction.

203 **Example 1.** The average is not an exactly mergeable summary

$$\bar{v}_A = \frac{1}{|A|} \sum_{X \in A} v(X)$$

	$v(A_1) = [2, 6]$	$\bar{v}_{A_1}(v) = 4$	$v(A_2) = [3, 4, 5]$	$\bar{v}_{A_2} = 4$
204	$v(B_1) = [1, 3, 5]$	$\bar{v}_{B_1}(v) = 3$	$v(B_2) = [1, 5]$	$\bar{v}_{B_2} = 3$
		$\bar{v}_{A_1 \cup B_1} = 3.4$		$\bar{v}_{A_2 \cup B_2} = 3.6$

205 **Example 2.** The median is not an exactly mergeable summary

$$\text{med}_A(v) = \text{sort}_A(v) \left[\left\lceil \frac{n_A}{2} \right\rceil \right]$$

	$v(A_1) = [3, 4, 1]$	$\text{med}_{A_1}(v) = 3$	$v(A_2) = [3, 8]$	$\text{med}_{A_2}(v) = 3$
206	$v(B_1) = [9, 6]$	$\text{med}_{B_1}(v) = 6$	$v(B_2) = [6, 2, 7]$	$\text{med}_{B_2}(v) = 6$
		$\text{med}_{A_1 \cup B_1}(v) = 4$		$\text{med}_{A_2 \cup B_2}(v) = 6$

207 **Example 3.** The 2nd is not an exactly mergeable summary

	$v(A_1) = [1, 3, 5]$	$2\text{nd}_{A_1}(v) = 3$	$v(A_2) = [3, 3, 6]$	$2\text{nd}_{A_2}(v) = 3$
208	$v(B_1) = [2, 5, 6]$	$2\text{nd}_{B_1}(v) = 5$	$v(B_2) = [4, 5, 7]$	$2\text{nd}_{B_2}(v) = 5$
		$2\text{nd}_{A_1 \cup B_1}(v) = 2$		$2\text{nd}_{A_2 \cup B_2}(v) = 3$

209 **Example 4.** The mode is not an exactly mergeable summary.

210 Let $A_x = \{a \in A : v(a) = x\}$ then $\text{mode}_A(v) \in \text{Argmax}_{x \in v(A)} |A_x|.$

	$v(A_1) = [x, x, x, y, y]$	$\text{mode}_{A_1}(v) = x$	$v(A_2) = [x, x, x, z, z]$	$\text{mode}_{A_2}(v) = x$
211	$v(B_1) = [y, y, y, z, z]$	$\text{mode}_{B_1}(v) = y$	$v(B_2) = [y, y, y, z, z]$	$\text{mode}_{B_2}(v) = y$
		$\text{mode}_{A_1 \cup B_1}(v) = y$		$\text{mode}_{A_2 \cup B_2}(v) = z$

212 Note that also $|A_{1x}| = |A_{2x}| = 3$ and $|B_{1y}| = |B_{2y}| = 3$, but $5 = |(A_1 \cup B_1)_y| \neq |(A_2 \cup B_2)_z| = 4$.

213 2.5 Conclusions

214 In measurement theory [18, 19] measurement scales are divided into absolute, ratio, interval, ordinal, and
215 nominal. The corresponding “best representatives” are count, geometric mean, average (arithmetic mean),
216 median, and mode. The count is an exactly mergeable summary (2.4.1.1). So are the geometric mean
217 (2.4.1.7) and the average (2.4.1.5), provided that we keep also the size of the corresponding set of units.

218 Median and mode are not exactly mergeable. A good exactly mergeable alternative is to use the cor-
219 responding frequency distribution (histogram or bar chart). In the case of a large number of categories,
220 less frequent categories can be combined into a common category. By using the frequency distribution also
221 for aggregation of numerical (ratio and interval) variables, we get a uniform representation for all types of
222 variables.

223 3 Discovering patterns and trends with interval-valued data

224 P. Brito, A.P. Duarte Silva

225 3.1 Context

226 This study concerns the Portuguese Labour Force Survey (LFS), analysing data from the 1st trimester of
227 2008 and the 4th trimester of 2010. We only consider people who were unemployed at the time of the survey
228 (had no job and were looking for one), and focus on the Activity Time (in years)(AT) and Unemployment
229 Time (in months) (UT). Disregarding records with missing values, and keeping only those from mainland
230 Portugal (i.e. excluding Madeira and Azores), we end up with 1150 observations in 2008 and 1569 in 2010.

231 These micro-data were then gathered, in each case, on the basis of Gender (Mas, Fem), Region (North,
232 Centre, Lisbon and Tagus Valley (LTV), South), Age-Group (Young: 15-24, Prime: 25-44, Mature: 45 and
233 above) and Education (Basic or less, Secondary, Higher), leading to 58 sociological groups in 2008 (T1) and
234 68 in 2010 (T4) (as some of the 72 possible combinations do not occur) and which constitute the statistical
235 units to be analysed.

236 We note that although the individuals at micro data level are not the same in 2008 and 2010, the
237 aggregate units formed correspond to the same sub-populations - e.g., Young Women, from the North, with
238 Secondary Education - and data are hence comparable at aggregate level.

239 The objective of this study is to cluster the aggregate units in each year, and compare the obtained
240 partitions, trying to get insights about the dynamics between 2008 and 2010. For this purpose, we rely on the
241 parametric model for interval-valued variables proposed in [20] and the model-based clustering methodology
242 developed in [21]. Data aggregation as well as all analysis are done with R package MAINT.Data [22, 23].

243 3.2 LFS Interval data

244 Let $S = \{s_1, \dots, s_n\}$, be the set of n units under analysis. An interval-valued variable is defined by an
245 application

$$246 Y : S \rightarrow B \text{ such that } s_i \rightarrow Y(s_i) = [l_i, u_i]$$

247 where B is the set of all intervals of an underlying set $O \subseteq \mathbb{R}$.

248 Let I be an $n \times p$ data array representing the values of p interval-valued variables on S . Each row of I
 249 corresponds to an element $s_i \in S$, represented by a p -dimensional vector of intervals, $I_i = (I_{i1}, \dots, I_{ip}) =$
 250 $([l_{i1}, u_{i1}], \dots, [l_{ip}, u_{ip}]), i = 1, \dots, n$, as in Table 1.

Table 1: Array of interval-valued data

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

251 In our case, for each aggregate unit, in each year, the minimum and maximum values of each of the
 252 Activity Time (AT) and Unemployment Time (UT) were recorded. As a result, each group is described by
 253 two intervals, that represent the within range of variation of the Activity Time and Unemployment Time in
 254 the corresponding year. Table 2 displays some rows of the 2008 and 2010 data arrays.

Table 2: Data in 2008 (left) and 2010 (right), partial views.

Unit	AT	UT	Unit	AT	UT
Fem-Cen-Mat-Bas	[23.0, 47.0]	[1.0, 108.0]	Fem-Cen-Mat-Bas	[7.0, 51.0]	[1.0, 131.0]
Fem-Cen-Pri-Bas	[2.0, 34.0]	[2.0, 147.0]	Fem-Cen-Pri-Bas	[4.0, 31.0]	[1.0, 130.0]
\dots	\dots	\dots	\dots	\dots	\dots
Mas-Cen-You-Sec	[0.0, 4.0]	[2.0, 3.0]	Mas-Cen-You-Sec	[2.0, 5.0]	[2.0, 23.0]

255 Comparing the interval data arrays for 2008 and 2010, we could observe that in several cases the UT
 256 interval became much wider within the two years, with the maximum value for UT showing a large increase.
 257 This is specially the case for groups with higher education levels, being not so frequent and clear in groups
 258 with basic education. Examples of such cases are the Mature Women from the South with Secondary
 259 education (Fem-South-Mat-Sec), for whom the UT interval went from [3, 26] to [1, 131], or Prime Man from
 260 the Centre with Superior education (Mas-Cen-Pri-Sup), for whom it changed from [7, 13] to [1, 171].

261 Figure 2 displays the interval-valued data separately for the two years under analysis and the three
 262 education levels. We note that in 2008 the UT intervals (along the horizontal axis) differ considerably across
 263 education levels, getting larger as education level decreases. However, these intervals are much wider in
 264 2010 than in 2008 for groups with superior education (upper figures), somehow for groups with secondary
 265 education (middle figures), but not so much for groups with basic education (lower figures). This effect
 266 reduces the differences in the UT intervals across education levels observed in 2008. In both years, the
 267 Activity Time variability increases as education level decreases (intervals along the vertical axis get wider).
 268 However, we do not observe remarkable changes in AT from 2008 to 2010, for any of the three education
 269 levels.

270 3.3 Model

271 The value of an interval-valued variable Y_j for each $s_i \in S$ is usually defined by the lower and upper bounds
 272 l_{ij} and u_{ij} of $I_{ij} = Y_j(s_i)$. For modelling purposes, however, we consider an alternative parameterisation,
 273 representing each interval $Y_j(s_i)$ by the MidPoint $c_{ij} = \frac{l_{ij} + u_{ij}}{2}$ and Range $r_{ij} = u_{ij} - l_{ij}$ of I_{ij} .
 274 The Gaussian model (see [20]) assumes a multivariate Normal distribution for the MidPoints C and the logs

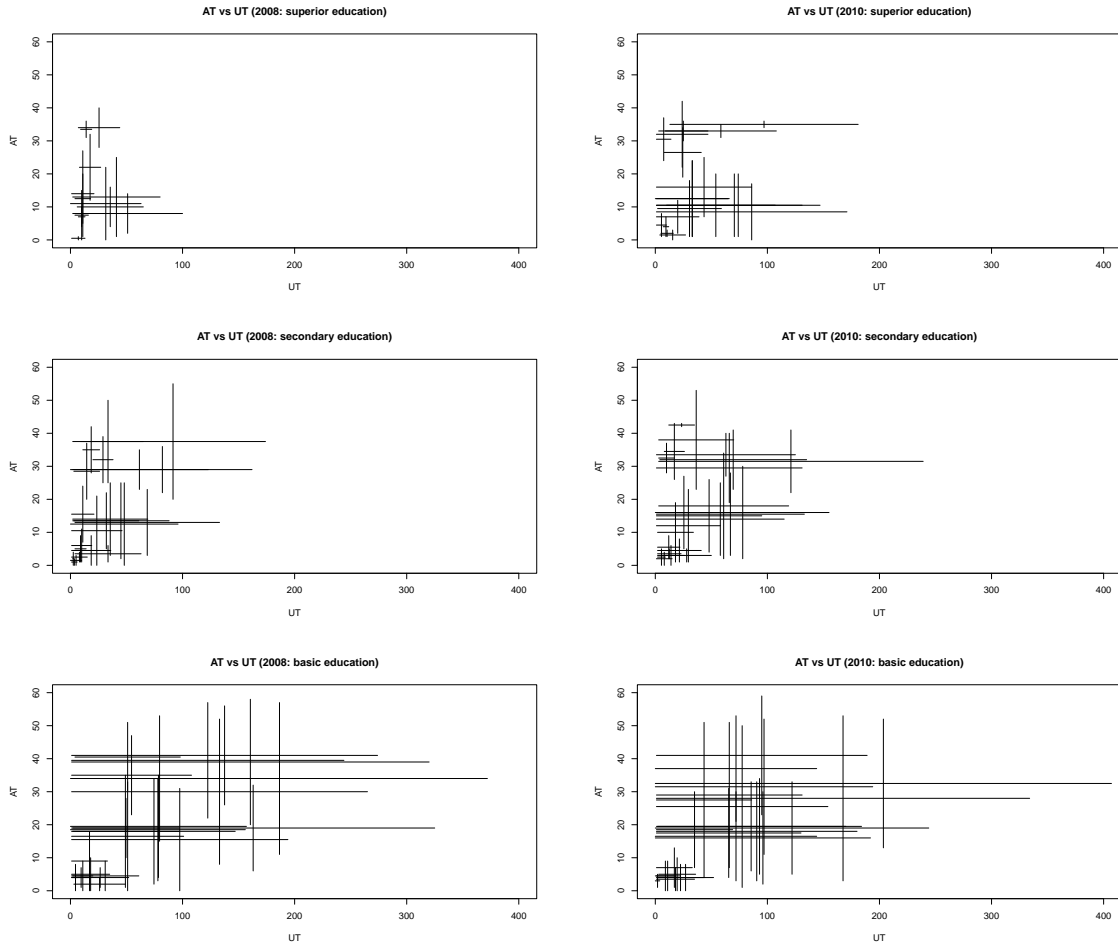


Figure 2: Interval data for 2008 (left) and 2010 (right), groups with superior education (top), secondary education (center), and basic education (bottom).

275 of the Ranges, $R^* = \ln(R)$, $(C, R^*) \sim \mathcal{N}_{2p}(\mu, \Sigma)$, with $\mu = [\mu_C^t, \mu_{R^*}^t]^t$ and $\Sigma = \begin{pmatrix} \Sigma_{CC} & \Sigma_{CR^*} \\ \Sigma_{R^*C} & \Sigma_{R^*R^*} \end{pmatrix}$ where
 276 μ_C and μ_{R^*} are p -dimensional column vectors of the mean values of, respectively, the MidPoints and the
 277 Log-Ranges, and $\Sigma_{CC}, \Sigma_{CR^*}, \Sigma_{R^*C}$ and $\Sigma_{R^*R^*}$ are $p \times p$ matrices with their variances and covariances.

278 We note that the model does not allow considering observations with degenerate intervals, where the
 279 range is null.

280 This model allows for the application of classical inference methods; however one should keep in mind
 281 that the MidPoint c_{ij} and the Range r_{ij} of the value of an interval-valued variable $I_{ij} = Y_j(s_i)$ pertain to one
 282 same variable, and must therefore be considered together. Specific configurations of the global covariance
 283 matrix allow taking into account the link that may exist between MidPoints and Ranges of the same or
 284 different variables. We consider the following configurations:

- 285 *C1* - Non-restricted configuration: allowing for non-zero correlations among all MidPoints and Log-Ranges;
- 286 *C2* - Interval-valued variables Y_j are independent, but for each variable, the MidPoint may be correlated
 287 with its Log-Range: $\Sigma_{CC}, \Sigma_{CR^*} = \Sigma_{R^*C}, \Sigma_{R^*R^*}$ all diagonal;
- 288 *C3* - MidPoints (Log-Ranges) of different variables may be correlated, but no correlation between MidPoints
 289 and Log-Ranges is allowed: $\Sigma_{CR^*} = \Sigma_{R^*C} = \mathbf{0}$;
- 290 *C4* - All MidPoints and Log-Ranges are uncorrelated, both among themselves and between each other: Σ
 291 diagonal.

292 From the Normality assumption it obviously follows that imposing non-correlations with Log-Ranges is
 293 equivalent to imposing non-correlations with Ranges. In cases *C2*, *C3* and *C4*, Σ may be written as a block
 294 diagonal matrix, after a possible rearrangement of rows and columns.

295 The mean vector and the variance-covariance matrix may be estimated by maximum likelihood. In the
 296 restricted configurations *C2*, *C3*, and *C4*, estimation may be done block-wise (see [20]).

297 3.4 Model-based Clustering

298 Model-based Clustering considers the data as coming from a distribution that is a mixture of several com-
 299 ponents [24, 25, 26]. Each component is then associated with a cluster, characterized by a conditional
 300 density/mass function, and has a probability or “weight”. When the conditional probability is specified as
 301 the multivariate Gaussian, the model will be a finite mixture of multivariate Normals, known as the Gaussian
 302 mixture model.

303 The model parameters for each component, and the membership (posterior) probabilities of each unit,
 304 must be estimated, this is commonly accomplished by the Expectation-Maximisation (EM) algorithm [27].
 305 This algorithm alternates an expectation (E) step, where the expectation of the log-likelihood at the cur-
 306 rent parameter estimates is computed, and a maximisation (M) step, where parameters are estimated by
 307 maximising the expected log-likelihood found in the E step.

308 Model-based Clustering of interval-valued data has been developed in [21], considering the Gaussian
 309 model described above (see Section 3.3), where the EM algorithm has been suitably adapted to the likelihood
 310 maximisation for the different covariance configurations.

311 The finite mixture model with k components for a data vector \mathbf{x} is defined as

$$f(\mathbf{x}; \Gamma) = \sum_{\ell=1}^k \pi_{\ell} f_{\ell}(\mathbf{x}; \theta_{\ell}) \quad (2)$$

312 where all weights π_ℓ are positive and $\pi_1 + \dots + \pi_k = 1$; θ_ℓ are the parameters of the conditional distribution
 313 of component ℓ . In the Gaussian case, we consider two alternatives: a homoscedastic setup, where the
 314 covariance matrix is constant across components, and the conditional distribution is given by $N(\mu_\ell, \Sigma)$, and
 315 a heteroscedastic setup with one covariance matrix per component, with conditional distribution $N(\mu_\ell, \Sigma_\ell)$.
 316 Maximum likelihood parameter estimation involves the maximisation of the log-likelihood function

$$\ln L(\Gamma) = \sum_{i=1}^n \ln f(\mathbf{x}_i; \Gamma) \quad (3)$$

317 with $\Gamma = (\pi_1, \dots, \pi_k, \mu_1, \dots, \mu_k, \Sigma)$ in the homoscedastic case, and $\Gamma = (\pi_1, \dots, \pi_k, \mu_1, \dots, \mu_k, \Sigma_1, \dots, \Sigma_k)$
 318 in the heteroscedastic case.

319 In Model-based Clustering of interval data, $\mathbf{x}_i = [\mathbf{c}_i^t, \mathbf{r}_i^{*t}]^t$ is defined as the $2p$ -dimensional vector
 320 comprising all the MidPoints and Log-Ranges for s_i . In the unrestricted case, the M-step formulas for $\hat{\Sigma}$ or
 321 $\hat{\Sigma}_\ell$ are the classical ones; for the restricted configurations, $\hat{\Sigma}$ or $\hat{\Sigma}_\ell$ are obtained maximising the likelihood
 322 for each block separately (see [20]).

323 The covariance configuration and the number of components k are selected as those that minimise the
 324 Bayesian Information Criterion (BIC) [28].

325 3.5 Clustering of the LFS data

326 The method presented above was applied to the LFS data described in Section 3.1, separately for 2008 and
 327 2010. However, we had to disregard six units in 2008 and two in 2010, since they presented a degenerate
 328 interval in at least one of the two variables. The units removed in 2008 were: F-Center-Young-Sup, F-LTV-
 329 Young-Sup, F-North-Young-Sup, F-South-Mature-Sup, M-Center-Prime-Sec, M-Center-Young-Sup, and in
 330 2010: F-Center-Young-Sec and M-North-Young-Sup.

331 Figure 3 shows the BIC values for $k = 2, \dots, 8$, all four covariance configurations, and both homoscedastic
 332 and heteroscedastic setups.

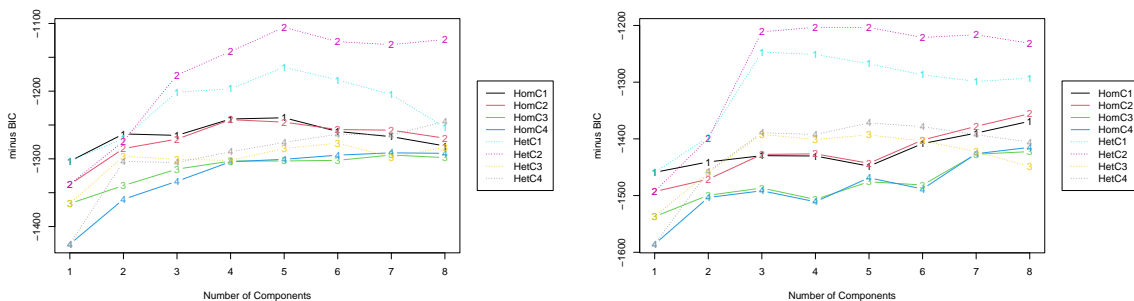


Figure 3: BIC values for 2008 (left) and 2010 (right).

333 We notice that the lowest BIC value is attained for configuration $C2$ and a heteroscedastic setup in both
 334 cases, with $k = 5$ in 2008 and $k = 4$ in 2010.

335 The best partition obtained for 2008 is displayed in Table 3, and Table 4 gathers the mean values per
 336 component for the four indicators. Figure 4 displays the parallel coordinate plot for this partition, providing
 337 some insights for its characterisation.

Table 3: Best partition in 2008.

Component	Composition
CP1	Fem-Cen-Mat-Bas ; Fem-LTV-Mat-Sec ; Fem-LTV-Mat-Sup ; Fem-Nor-Mat-Sec ; Fem-South-Mat-Sec ; Mas-Cen-Mat-Bas ; Mas-Cen-Mat-Sec ; Mas-LTV-Mat-Sec ; Mas-LTV-Mat-Sup ; Mas-Nor-Mat-Bas ; Mas-Nor-Mat-Sec ; Mas-Nor-Mat-Sup ; Mas-South-Mat-Sec
CP2	Fem-Cen-Pri-Bas ; Fem-LTV-Pri-Bas ; Fem-South-Pri-Bas ; Mas-Cen-Pri-Bas ; Mas-LTV-Pri-Bas ; Mas-LTV-Pri-Sec ; Mas-Nor-Pri-Bas ; Mas-South-Pri-Bas
CP3	Fem-Cen-Pri-Sup ; Fem-Cen-You-Sec ; Fem-LTV-You-Bas ; Fem-LTV-You-Sec ; Fem-Nor-You-Sec ; Fem-South-Pri-Sup ; Fem-South-You-Sup ; Mas-Cen-Pri-Sup ; Mas-Cen-You-Bas ; Mas-Cen-You-Sec ; Mas-LTV-You-Bas ; Mas-LTV-You-Sec ; Mas-Nor-You-Sec ; Mas-South-Pri-Sec ; Mas-South-Pri-Sup
CP4	Fem-Cen-Pri-Sec ; Fem-Cen-You-Bas ; Fem-LTV-Pri-Sec ; Fem-LTV-Pri-Sup ; Fem-Nor-Pri-Sec ; Fem-Nor-Pri-Sup ; Fem-Nor-You-Bas ; Fem-South-Pri-Sec ; Fem-South-You-Bas ; Fem-South-You-Sec ; Mas-LTV-Pri-Sup ; Mas-Nor-Pri-Sec ; Mas-Nor-Pri-Sup ; Mas-Nor-You-Bas ; Mas-South-You-Bas ; Mas-South-You-Sec
CP5	Fem-LTV-Mat-Bas ; Fem-Nor-Mat-Bas ; Fem-Nor-Pri-Bas ; Fem-South-Mat-Bas ; Mas-LTV-Mat-Bas ; Mas-South-Mat-Bas

Table 4: Mean values by component in 2008.

Indicator	CP1	CP2	CP3	CP4	CP5
AT.MidP	32.88	16.93	5.83	8.52	33.75
AT.LogR	2.84	3.33	1.87	2.54	3.58
UT.MidP	44.04	66.93	8.20	31.54	150.50
UT.LogR	3.92	4.84	2.26	4.03	5.69

338 Component 5 comprises Mature units with Basic education, and is characterised by high Activity and
 339 Unemployment Times, both with high variability (as conveyed by the log-ranges).

340 Component 1 is also composed by Mature units, now mostly with some education, with high Activity
 341 Times, lower Unemployment Times, both with not so high variability.

342 Component 3 is formed by Young and a few Prime units, it is characterised by low Activity and Unem-
 343 ployment Times, both with low variability.

344 Component 4 is mostly composed by Prime units with some education, and a few Young units. Its
 345 characterisation is similar to that of Component 3, but less pronounced; however, Unemployment time
 346 shows higher variability.

347 Finally, Component 2 comprises Prime units mostly with Basic education; it shows intermediate Activity
 348 and Unemployment Times, both with relatively high variability.

349

350 Table 5 describes the best partition obtained for 2010, Table 6 provides the corresponding mean values,
 351 and Figure 5 displays the parallel coordinate plot.

352 Component 4 gathers Young units and is characterised by low Activity and Unemployment Times, both
 353 with low variability. It somehow corresponds to Component 3 of 2008, although now with no Prime units.
 354 In comparison, and as expected, Activity Time is on average lower and with lower variability. However,
 355 Unemployment Time is now on average higher, and with higher variability, which may be a consequence of
 356 the European sovereign debt crisis.

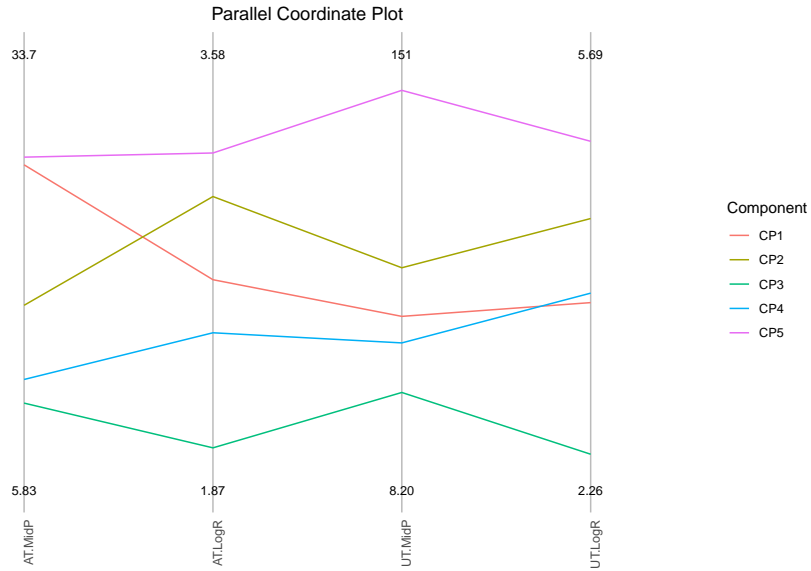


Figure 4: Parallel coordinate plot for 2008, best model.

357 Component 1 is essentially formed by Mature units, irrespective of education level, coming from CP1 and
 358 CP5 of the 2008 partition. Activity and Unemployment Times are high, as expected, both with relatively
 359 high variability.

360 Component 2 is mostly composed by Prime units, with a few Mature. It is characterised by the high
 361 variability of both Activity and Unemployment Times, and high Unemployment Time.

362 Finally, Component 3, similarly to Component 4 of 2008, gathers mainly Prime and some Young units,
 363 with intermediate Activity and Unemployment Times, and respective variabilities.

364

365 We note that, unlike what happened in 2008, in 2010 education seems not to play a role in the formation
 366 of clusters. In particular, in 2008 there is a clear separation of Mature groups in two clusters, mainly
 367 distinguished in terms of education, while in the 2010 partition most Mature are grouped in one single
 368 cluster.

369 As a final remark, we observe that, although the individuals at micro data level are not the same in 2008
 370 and 2010, the aggregate units formed correspond to the same sub-populations, allowing for a comparative
 371 analysis of the resulting partitions.

372 In both years, given the two variables used, age seems to be a driving force in the formation of clusters;
 373 in 2008 education also appears to play a role, but not so much in 2010.

374 It is noteworthy that, although some cluster correspondences may be identified, the partition is essentially
 375 not stable from 2008 to 2010, the sociological units gather in a different way. This may be the result of the
 376 change in the Portuguese labour market created by the economic crisis.

Table 5: Best partition in 2010.

Component	Composition
CP1	Fem-Cen-Mat-Sec ; Fem-LTV-Mat-Sec ; Fem-LTV-Mat-Sup ; Fem-Nor-Mat-Bas ; Fem-Nor-Mat-Sec ; Fem-Nor-Mat-Sup ; Fem-South-Mat-Bas Fem-South-Mat-Sec ; Fem-South-Mat-Sup ; Fem-South-Pri-Bas ; Mas-Cen-Mat-Bas ; Mas-Cen-Mat-Sec ; Mas-LTV-Mat-Sec ; Mas-LTV-Mat-Sup Mas-Nor-Mat-Bas ; Mas-Nor-Mat-Sec ; Mas-Nor-Mat-Sup ; Mas-South-Mat-Bas ; Mas-South-Mat-Sec ; Mas-South-Mat-Sup
CP2	Fem-Cen-Mat-Bas ; Fem-Cen-Mat-Sup ; Fem-Cen-Pri-Bas ; Fem-Cen-Pri-Sup ; Fem-LTV-Mat-Bas ; Fem-LTV-Pri-Bas ; Fem-LTV-Pri-Sec ; Fem-LTV-Pri-Sup ; Fem-Nor-Pri-Bas ; Fem-Nor-Pri-Sec ; Fem-South-Pri-Sec ; Mas-Cen-Pri-Sup ; Mas-LTV-Mat-Bas ; Mas-LTV-Pri-Bas ; Mas-Nor-Pri-Bas ; Mas-Nor-Pri-Sec ; Mas-South-Pri-Bas ; Mas-South-Pri-Sec ; Mas-South-Pri-Sup
CP3	Fem-Cen-Pri-Sec ; Fem-Nor-Pri-Sup ; Fem-Nor-You-Bas ; Fem-South-Pri-Sup ; Fem-South-You-Bas ; Fem-South-You-Sec ; Mas-Cen-Pri-Bas ; Mas-Cen-Pri-Sec ; Mas-LTV-Pri-Sec ; Mas-LTV-Pri-Sup ; Mas-Nor-Pri-Sup ; Mas-Nor-You-Bas ; Mas-South-You-Bas
CP4	Fem-Cen-You-Bas ; Fem-Cen-You-Sup ; Fem-LTV-You-Bas ; Fem-LTV-You-Sec ; Fem-LTV-You-Sup ; Fem-Nor-You-Sec ; Fem-South-You-Sup ; Mas-Cen-You-Bas ; Mas-Cen-You-Sec ; Mas-LTV-You-Bas ; Mas-LTV-You-Sec ; Mas-LTV-You-Sup ; Mas-Nor-You-Sec ; Mas-South-You-Sec

Table 6: Mean values by component in 2010.

Indicator	CP1	CP2	CP3	CP4
AT.MidP	32.63	17.01	9.56	3.46
AT.LogR	2.69	3.28	2.66	1.64
UT.MidP	69.88	70.09	25.62	11.94
UT.LogR	4.45	4.89	3.85	2.73

4 An illustration of the use of the measures s-concordance and s-discordance in applications

S. Korenjak-Černe, J. Dobša, E. Diday

A “similarity” as a “concordance” in data analysis represents a mathematical modeling of the words “similarity” and “concordance” used in our natural language. The similarity measure quantifies the similarity between two objects and has a symmetric property, while the concordance measures the similarity between an object and (with) a collection of objects and therefore has no symmetric property. Thus, similarity and concordance express two different kinds of knowledge.

The concordance measure based on symbolic data description was introduced by Diday in 2020 [29]. It is named s-concordance with the prefix “s” because it is defined for symbolic data where the objects represent aggregations of individuals, i.e., classes, so the definition falls within the framework of symbolic data analysis (SDA) [29]. A class has high concordance with a given collection of classes for a category x if that category is frequent in that class and if, in addition, there are numerous classes in the given collection of classes for which category x is also frequent.

The definition of s-concordance is based on two functions:

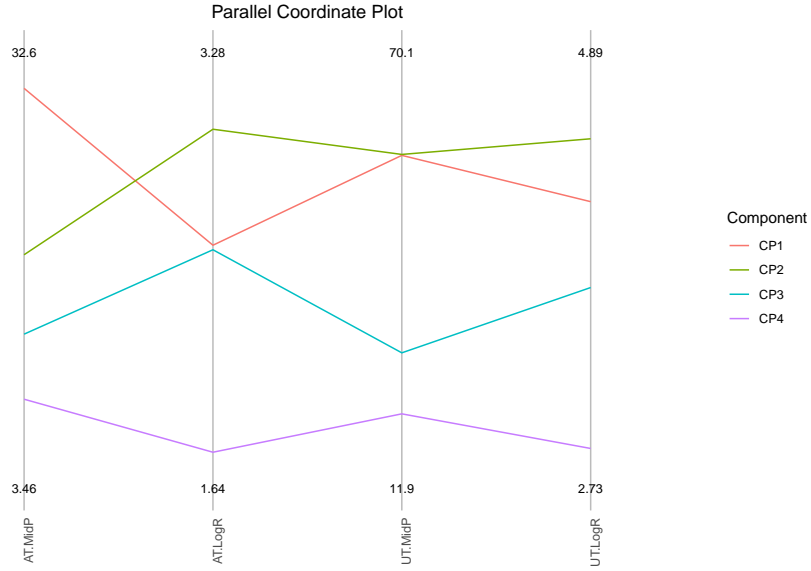


Figure 5: Parallel coordinate plot for 2010, best model.

- 392 • $f_c(x)$ is the proportion of individuals with category x in the class c and therefore measures the fit
393 between category x and the class c ,
- 394 • $g_x(c, P)$ is the proportion of such classes c' from the collection P , for which the frequency $f_{c'}(x)$ is
395 close to the frequency $f_c(x)$. Here, by the term “is close to” we mean “to be on the same subinterval
396 of the interval $[0, 1]$ ”.

397 An axiomatic definition of s -concordance was given by Diday ([29], [30]) where he presented examples
398 of concordance and related discordance measures. In our illustrative examples, we have focused on three of
399 them:

- 400 1. $S_{conc1}(c, P; x) = g_x(c, P)$
401 The higher the proportion of classes c' with values $f_{c'}(x)$ on the same subinterval as $f_c(x)$, the higher
402 the s -concordance of class c with collection P .
- 403 2. $S_{conc2}(c, P; x) = f_c(x) \cdot g_x(c, P)$
404 A higher proportion of individuals with category x within class c (i.e., high $f_c(x)$) increases the
405 s -concordance of class c with collection P by the same proportion of classes.
- 406 3. $S_{disc3}(c, P; x) = \frac{f_c(x)}{1 + g_x(c, P)}$
407 A high proportion of individuals with category x within class c (i.e., high $f_c(x)$) given a low proportion
408 of classes with the same or similar values (i.e., a low $g_x(c, P)$) means that the value x is in some way
409 characteristic of class c .

410 Note that none of the pairs of these concordance and discordance measures are complementary opposites.
411 Two very different examples of the use of the presented measures are presented below.

412 **4.1 Application on the data of international measurement of reading**
 413 **achievement among young students**

414 In the first example, we are interested in the concordances and discordances of countries based on the
 415 proportion of students who scored at or above a high (high or advanced) level on the traditional paper
 416 reading assessment in the Progress in International Reading Literacy Study (PIRLS [31]). The PIRLS is
 417 an international assessment and research project designed to measure the reading achievement of fourth
 418 graders and the instructional practices of schools and teachers. The reading achievement scale is derived
 419 from several variables that measure the quality of reading. More detailed information can be found on the
 420 website of the cited reference (PIRLS 2016 User Guide, Chapter 4, p. 63). We focus our study on the 2016
 421 survey data from fifty participating countries.

422 In this case, individuals are students and classes are countries. The collection P is the set of all 50
 423 countries participating in the study. The x category includes high and advanced levels of traditional reading
 424 achievement. For each country c , $f_c(x)$ indicates the proportion of students who achieved a high or advanced
 425 level of traditional reading achievement. To obtain the function $g_x(c, P)$, we divided the interval $[0, 1]$ into
 426 five equidistant subintervals and, based on the values $f_{c'}(x)$, $c' \in P$, obtained the distribution of countries
 427 shown in Table 7. Figure 6 presents the positioning of the countries in the plane with axes $f_c(x)$ and $g_x(c, P)$.

Table 7: Distribution of countries by proportion of students with high or advanced levels of tradi-
 tional reading achievement in the PIRLS 2016 survey

Subinterval	Number of countries	Proportion
$[0, 0.2]$	12	0.24
$(0.2, 0.4]$	8	0.16
$(0.4, 0.6]$	23	0.46
$(0.6, 0.8]$	7	0.14
$(0.8, 1]$	0	0
Sum	50	1

428 As it can be seen from Table 7, almost half of the countries have more than 40% and up to 60% of students
 429 achieving at least a high level of reading literacy on paper, so we can say that each of these countries is
 430 concordant with the collection of all countries. This interpretation is captured by the measure S_{conc1} . In
 431 Figure 6, these countries are positioned in the upper right place.

432 The second chosen s-concordant measure S_{conc2} can be interpreted as the area of the rectangles in the
 433 plane in Figure 6, for each country c determined by the values $f_c(x)$ and $g_x(c, P)$. The larger area of the
 434 rectangle corresponding to the country expresses a higher concordance of this country with the collection
 435 of all countries. Among the 23 countries with a proportion of well-skilled students between 0.4 and 0.6,
 436 Chinese Taipei had the highest proportion, and the value of its S_{conc2} is the largest for this country. Thus,
 437 taking into account both a good fit between the country and observed literacy achievement (in our case high
 438 or advanced literacy) which results in high $f_c(x)$, and the frequent occurrence of such fits among countries,
 439 we can say that Chinese Taipei is the country for which the concordance with a collection of all countries is
 440 the highest.

441 There are seven countries where the proportion of well-skilled students is above 0.6 (positioned at the
 442 bottom right of Figure 6). Based on their good fit with the category (high $f_c(x)$) and the rare frequency of
 443 such good fits, these most discordant countries can be identified by the highest s-discordance value S_{disc3} . In
 444 PIRLS 2016 data, the Russian Federation has the highest proportion of well-skilled students in paper reading.
 445 Since there are only 7 out of 50 countries in our data that have more than 60% of well-skilled students, the

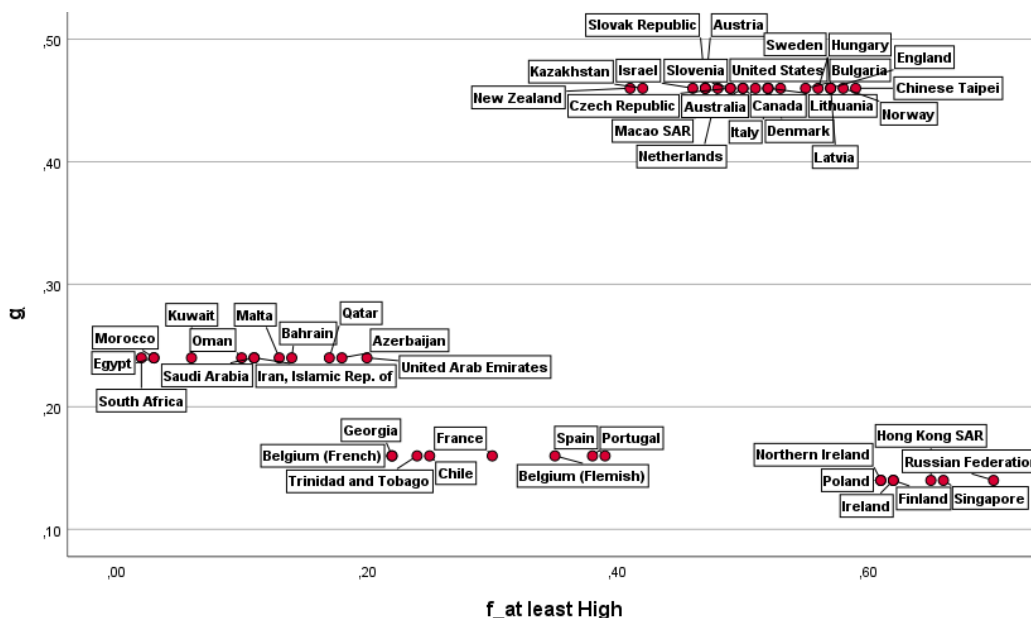


Figure 6: Positioning of the countries based on their values of the functions f and g for the traditional (paper) reading.

446 Russian Federation is the most discordant country in the collection of all participating countries.

447 4.2 Application on textual data with the comparison with Tf-Idf

448 In the second example, we explore the use of the discordance measure as an alternative to the well-known
 449 Tf-Idf (term frequency - inverse document frequency) measure in Text Mining. The vector space model for
 450 representing text collections of documents is represented by a term-document matrix in which the documents
 451 are represented by columns and the index terms used to index the document collection are represented by
 452 rows. The basic idea of Tf-Idf (see [32]) is to characterize a category (here presence of the term x) of a class
 453 of documents by its “relevance” compared to the other classes. The relevance of the term x to the class is
 454 high if

- 455 • the proportion of documents within the class containing that term is high and;
- 456 • the classes of the given partition P of the document collection in which it occurs are rare.

457 There are several variants of the Tf-Idf measure. We will use its basic form, which is for the term x and
 458 class of documents c defined as

$$459 \text{Tf-Idf}(x, c) = \frac{n(x, c)}{|c|} \cdot \frac{K}{k} = f_c(x) \cdot \frac{K}{k}$$

460 where

- $n(x, c)$ is the number of documents in the class c in which the term x occurs
- $|c|$ is the number of documents in the class c
- K is the total number of classes in the collection of documents
- 461 k is the number of classes in the collection of documents for which
 there is at least one document in which the term x occurs

462 Note that the definition of Tf-Idf does not take into account the differences between classes due to
 463 the number of documents in which the term occurs. In the definition of the s-discordance, however, these
 464 differences are included in the function $g_x(c, P)$

$$465 \quad S_{disc3}(c, P; x) = f_c(x) \cdot \frac{1}{1 + g_x(c, P)}$$

466 We illustrate the difference between these measures in the collection of 15 documents (book titles) from
 467 the field of data mining (DM documents), linear algebra (LA documents), and one document combining
 468 these two fields (Table 8) [33].

469

Table 8: Collection of textual documents

Doc.	Class	Text of the document
D1	DM	Survey of text mining: clustering, classification, and retrieval
D2	DM	Automatic text processing: the transformation analysis and retrieval of information by computer
D3	LA	Elementary linear algebra: A matrix approach
D4	LA	Matrix algebra and its applications in statistics and econometrics
D5	DM	Effective databases for text and document management
D6	DM, LA	Matrices, vector spaces, and information retrieval
D7	LA	Matrix analysis and applied linear algebra
D8	LA	Topological vector spaces and algebras
D9	DM	Information retrieval: data structures and algorithms
D10	LA	Vector spaces and algebras for chemistry and physics
D11	DM	Classification, clustering, and data analysis
D12	DM	Clustering of large data sets
D13	DM	Clustering algorithms
D14	DM	Document warehousing and text mining: techniques for improving business operations, marketing and sales
D15	DM	Data mining and knowledge discovery

470 The list of index terms consists of terms that appear in at least two documents, with so called
 471 stop words or words commonly used in a language sorted out, and word variants mapped in their
 472 base form. In this way, a list of 16 index terms was created.

473 The values of the function $g_x(c, P)$ are determined using five equidistant subintervals [0,0.2],
 474 (0.2,0.4], (0.4, 0.6], (0.6, 0.8], (0.8, 1.0]. In our case of two classes $g_x(c, P)$ can take only two
 475 values: 1, if $f_c(x)$ falls in the same interval for both classes; and 0.5, otherwise. In order to com-
 476 pare the values of $S_{disc3}(c, P; x)$ and $Tf-Idf(x, c)$, these values are normalized to the interval [0,1],
 477 i.e. divided by the maximum possible values of these functions in the case of two classes. Cal-
 478 culated values for both classes and for each of 16 terms are presented in Table 9. Identifying a
 479 term as characteristic for a class if it has a higher value of observed measures, we can see that
 480 both measures recognize relevant terms for classes of DM and LA documents. Values of measures
 481 $S_{disc3}(DM, P; x)$ and $Tf-Idf(x, DM)$ are greater for terms x related to the field of data mining
 482 (such as *classification, clustering, data, document, information, mining, retrieval, and text*) while
 483 measures $S_{disc3}(LA, P; x)$ and $Tf-Idf(x, LA)$ are greater for terms x related to the field of linear

484 algebra (*algebra, linear, matrix, space, and vector*).

485

Table 9: Values of Tf-Idf and s-discordance measures of index terms for classes DM (data mining documents) and LA (linear algebra documents). α -relevant values for $\alpha = 0.3$ are shown in bold. Terms which are recognized as relevant for classes by the s-discordance measure, and are not recognized as relevant by the Tf-Idf measure are shown framed, as well as their s-discordance and Tf-Idf measures for the respective class.

Term x	$S_{disc3}(DM, P; x)$	$Tf-Idf(x, DM)$	$S_{disc3}(LA, P; x)$	$Tf-Idf(x, LA)$
<i>algebra</i>	0.000	0.000	0.833	0.833
<i>algorithm</i>	0.150	0.200	0.000	0.000
<i>analysis</i>	0.150	0.100	0.125	0.083
<i>application</i>	0.000	0.000	0.333	0.333
<i>classification</i>	0.150	0.200	0.000	0.000
<i>clustering</i>	0.400	0.400	0.000	0.000
<i>data</i>	0.400	0.400	0.000	0.000
<i>document</i>	0.150	0.200	0.000	0.000
<i>information</i>	0.300	0.150	0.167	0.083
<i>linear</i>	0.000	0.000	0.333	0.333
<i>matrix</i>	0.000	0.000	0.500	0.500
<i>mining</i>	0.300	0.300	0.000	0.000
<i>retrieval</i>	0.400	0.200	0.167	0.083
<i>space</i>	0.100	0.050	0.500	0.250
<i>text</i>	0.400	0.400	0.000	0.000
<i>vector</i>	0.100	0.050	0.500	0.250

486 To distinguish relevant terms for classes based on defined measures of s-discordance and Tf-Idf
487 we define α -relevance for these measures. For a given threshold $\alpha > 0$, Tf-Idf is considered α -relevant
488 if it is at least α , and analogously, the S_{disc3} value is called α -relevant if it is at least α . By choosing
489 $\alpha = 0.3$, we show some differences between the measures in the recognition of relevant terms shown
490 in Table 9. In this case, the S_{disc3} measure recognized relevant terms more successfully than Tf-Idf,
491 because unlike Tf-Idf, S_{disc3} recognized the terms *information* and *retrieval* as relevant for the class
492 DM and also recognized the terms *space* and *vector* as relevant for the class LA. Reason for that
493 is that the Tf-Idf measure equalizes classes in which a specific term appears at least once, while
494 the s-discordance measure makes a difference according to the frequency of the term appearance in
495 classes.

496 4.3 Discussion and further work

497 We have shown two examples of possible applications of the new measures s-concordance and s-
498 discordance from different contexts.

499 In the first example, these measures were used to measure the concordance and discordance of a
500 single country with the collection of all countries. We used data from a PIRLS 2016 survey where we
501 focused on at least a high level of traditional paper reading assessment. Measures of s-concordance
502 and s-discordance are used to compare the country to the collection of all countries. Since in the
503 definitions of s-concordance and s-discordance we use the distribution of classes included with the

504 function g , which in application is usually based on the empirical distribution, a more detailed study
505 of its influence is needed in the future.

506 In the second example, we used the measure of s -discordance to identify relevant terms that are
507 characteristic of a particular class of documents. We compared our results with those obtained using
508 the Tf-Idf measure. In our case, the s -discordance measure identified more relevant terms for classes.

509 The Tf-Idf measure detects relevant terms for a class if that term occurs only in that class, while
510 the s -discordance measure detects relevant terms for a class if the frequency of their occurrence in
511 that class is greater than in other classes. Because of this property, the s -discordance measure could
512 be used to extend a weighting of terms when representing a document in a vector space model to
513 improve classification performance. It could also be used in sentiment analysis to automatically
514 capture a lexicon by calculation of s -discordance measure of index terms for classes of documents
515 with positive and negative sentiment.

516 5 Conclusion

517 In conclusion, this paper intends to draw attention to Symbolic Data Analysis (SDA) in the field of
518 Official Statistics, with a number of examples in the various sections that corroborate the potential of
519 the proposed methods. SDA has been a pioneering line of research in the treatment of unconventional
520 data, i.e. in the form of aggregated or, by their very nature, complex data. Nowadays, we refer to
521 the latter as data with a greater degree of granularity. The interest behind SDA is to be able to
522 extend statistical techniques and analysis of basic data to data representing classes of individuals.
523 These are typical data from Official Statistics which, for reasons of confidentiality, synthesis and
524 relative classification, are appropriately expressed in ranges of values and/or in divisions linked to
525 spatial classifications. The several SDA pilot contributions presented in this paper highlight the
526 application value and prospects that can be opened up for new developments in the extension of
527 analysis techniques to Big Data. They require, as exhibited, summarization methods and appropriate
528 data aggregation techniques that can be provided through symbolical data modelling. Each section
529 emphasises the importance of a contribution.

530 Section 2 presents new aggregation methods based on the notion of exactly unified summaries.
531 This property is generally not invoked in data reduction techniques and appropriate summaries and
532 representations. For this reason, it makes an original contribution to the creation of a method-
533 ological framework for the aggregation of data in the form of intervals, bar charts, and histograms.
534 Particularly interesting is the extension of the concept of unifiable summations to moments and
535 distributions by adding the dimension of the set of units. Finally, this first contribution shows how
536 the concept of unifiability can be extended to already synthesised or higher-level data.

537 The second contribution presented in this paper, in Section 3, focuses on an innovative approach
538 to identify patterns and trends by analysing interval data, with an application to data from the
539 Portuguese Labour Force Survey. This work shows how the representation of data in the form of
540 intervals allows summarising well the information expressed by aggregated numerical data when there
541 are no assumptions about the distribution within the given intervals of values. However, the strength
542 of the contribution is to consider the variability referring to the entire range of values between the

543 minimum and maximum, rather than just the central value as the centre of the distribution of values
544 within the range. The work presented in Section 3 focuses on a parametric model-based clustering
545 approach for aggregate interval-valued data. The application strength of the proposed methodology
546 is demonstrated on data from the Portuguese Labour Force Survey - a typical study provided by
547 NSI's - for which it is proposed to cluster sociological units described by the corresponding range
548 of Activity Time and Unemployment Time. Although the individuals at the micro-data level are
549 not the same in 2008 and 2010, the aggregate units formed correspond to the same sub-populations,
550 allowing for a comparative analysis of the resulting partitions. The proposed methodology also made
551 it possible to assess the major influence of the analysis variables, among which age plays a more
552 prevalent role than education.

553 The third contribution refers to one of the latest works that Edwin Diday was developing with the
554 section co-authors on s-concordance and s-discordance measures. These are intended to express the
555 degree of similarity between a class of individuals and a collection of classes based on the frequency of
556 the categories used to describe them. Two applications highlight the value of the proposed measures
557 and the applicative contribution. The first concerns data from the international measurement of
558 reading achievement among young students in the Progress in International Reading Literacy Study
559 PIRLS 2016 survey in fifty participating countries. The proposed measures are used to assess the
560 concordance and discordance of a single country with the collection of all countries. A few countries
561 with the highest percentage of students who are well qualified in reading and writing are identified
562 as the countries with the highest s-discordance values, as they are very different from most other
563 countries in this respect. The second application, on a collection of textual documents, highlights
564 the effectiveness of using the discordance measure to identify relevant terms that are characteristic
565 of a particular class of documents, as an alternative to the usual term frequency measure - inverse
566 document frequency (Tf-Idf) in Text Mining. As main result of such application, the novel measures
567 have allowed recognizing relevant terms more successfully than Tf-Idf.

568 SDA is still a flourishing and exciting field of research, with obvious potential in the context of
569 Official Statistics, both for analysing aggregate data and for providing tools for constructing com-
570 posite indicators that take into account the distribution of observed phenomena over time and space.
571 We are therefore confident that attention may be focused on the ongoing and future developments
572 of SDA. The field of Official Statistics remains particularly relevant to take advantage of the appli-
573 cation of these techniques and to suggest new approaches. A collaboration with researchers from
574 the NSI's is desirable to demonstrate the explanatory power of the SDA approach, also compared
575 with modern Machine Learning techniques that often do not guarantee equal interpretative ability.

576 Acknowledgments

577 **Special acknowledgments go to Edwin Diday who, with his passion and focus on the**
578 **research that has crowned his entire life, honored the NTTS2023 conference session**
579 **and the event as a whole with his presence.**

580 Section 2 by V. Batagelj is an elaboration of ideas presented at the 7th Workshop on Symbolic
581 Data Analysis, SDA 2018, held in Viana do Castelo, Portugal, 18-20 October 2018. It was presented

582 at NTTS2023 – Conference on New Techniques and Technologies for Statistics, 6-10 March 2023
583 (Brussels, Belgium). This work is supported in part by the Slovenian Research Agency (research
584 program P1-0294, research program CogniCom (0013103) at the University of Primorska, and re-
585 search projects J5-2557, J1-2481, and J5-4596), and prepared within the framework of the COST
586 action CA21163 (HiTEc).

587 The work of P. Brito and A.P. Duarte Silva (presented in Section 3) was supported by National
588 Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia,
589 within projects LA/P/0063/2020, DOI 10.54499/LA/P/0063/2020; and UID/GES/00731/2019.

590 Section 4 by S. Korenjak-Černe, J. Debša, and E. Diday is based on the work *An illustration of*
591 *the use of the measures s-concordance and s-discordance in applications* presented to the Conference
592 on New Techniques and Technologies for Statistics NTTS2023, 6-10 March 2023, Brussels, Belgium.
593 This work is supported in part by the Slovenian Research Agency (research program P1-0294).

594 References

- 595 [1] Bock HH, Diday E. Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical
596 Information from Complex Data. Springer Verlag; 2000.
- 597 [2] Diday E. The symbolic approach in clustering and related methods of Data Analysis. In: H BH,
598 editor. Classification and Related Methods of Data Analysis, Proceedings of the First Conference
599 of the International Federation of Classification Societies (IFCS-87): Technical University of
600 Aachen. North-Holland; 1988. p. 673-84.
- 601 [3] Diday E. Introduction à l'approche symbolique en Analyse des Données. In: Actes des Premières
602 Journées Symbolique-Numérique. Université Paris IX Dauphine; 1987. p. 21-56.
- 603 [4] Diday E. Introduction à l'approche symbolique en analyse des données. RAIRO - Operations
604 Research. 1989;23(2):193-236.
- 605 [5] Diday E, Noirhomme-Fraiture M. Symbolic Data Analysis and the SODAS Software. John
606 Wiley & Sons, Ltd; 2008.
- 607 [6] Kejžar N, Korenjak-Černe S, Batagelj V. Clustering of modal-valued symbolic data. Advances
608 in Data Analysis and Classification. 2021;15(2):513-41.
- 609 [7] Beliakov G, Pradera A, Calvo T. Aggregation Functions: A Guide for Practitioners. Studies in
610 Fuzziness and Soft Computing. Springer Berlin Heidelberg; 2007.
- 611 [8] Torra V, Narukawa Y. Modeling Decisions: Information Fusion and Aggregation Operators.
612 Cognitive Technologies. Springer Berlin Heidelberg; 2007.
- 613 [9] Grabisch M. Aggregation Functions. Encyclopedia of Mathematics and its Applications. Cam-
614 bridge University Press; 2009.

- 615 [10] Sola HB, Fernandez J, Mesiar R, Calvo T. Aggregation Functions in Theory and in Practice.
616 Proceedings of the 7th International Summer School on Aggregation Operators at the Public
617 University of Navarra, Pamplona, Spain, July 16-20, 2013. Advances in Intelligent Systems and
618 Computing. Springer Berlin Heidelberg; 2013.
- 619 [11] Halaš R, Gagolewski M, Mesiar R. New Trends in Aggregation Theory. Advances in Intelligent
620 Systems and Computing. Springer International Publishing; 2019.
- 621 [12] James S. An Introduction to Data Analysis using Aggregation Functions in R. Springer Inter-
622 national Publishing; 2016.
- 623 [13] Diday E. Probabilist, possibilist and belief objects for knowledge analysis. Annals of Operations
624 Research. 1995;55:225-76.
- 625 [14] Aitchison J. The Statistical Analysis of Compositional Data. Monographs on Statistics and
626 Applied Probability. Springer Netherlands; 1986.
- 627 [15] Ramsay J, Silverman BW. Functional Data Analysis. Springer Series in Statistics. Springer
628 New York; 2013.
- 629 [16] Rodriguez OR. RSDA 3.0.13: R to Symbolic Data Analysis; 2022. Available from: <https://cran.r-project.org/web/packages/RSDA/>.
- 631 [17] Agarwal PK, Cormode G, Huang Z, Phillips J, Wei Z, Yi K. Mergeable summaries. In: Krötzsch
632 M, editor. The 31st ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database
633 Systems (PODS '12), Proceedings. ACM, New York, NY, USA; 2012. p. 23-34.
- 634 [18] Stevens SS. On the theory of scales of measurement. Science. 1946;103(2684):677-80.
- 635 [19] Roberts FS. Measurement Theory: With Applications to Decisionmaking, Utility, and the
636 Social Sciences. vol. 7 of Encyclopedia of Mathematics and its Applications. Roberts FS, editor.
637 Cambridge University Press; 1985.
- 638 [20] Brito P, Duarte Silva AP. Modelling Interval Data with Normal and Skew-Normal Distributions.
639 Journal of Applied Statistics. 2012;39(1):3-20.
- 640 [21] Brito P, Duarte Silva AP, Dias JG. Probabilistic Clustering of Interval Data. Intelligent Data
641 Analysis. 2015;19(2):293-313.
- 642 [22] Duarte Silva AP, Brito P, Filzmoser P, Dias JG. MAINT.Data: Modelling and analysing interval
643 data in R. The R Journal. 2021;13(2):336-64.
- 644 [23] Duarte Silva AP, Brito P. MAINT.Data: Model and Analyse Interval Data; 2023. R package
645 version 2.7.1. Available from: <https://CRAN.R-project.org/package=MAINT.Data>.
- 646 [24] Banfield JD, Raftery AE. Model-based Gaussian and non-Gaussian clustering. Biometrics.
647 1993:803-21.

- 648 [25] Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation.
649 Journal of the American Statistical Association. 2002;97(458):611-31.
- 650 [26] McLachlan GJ, Peel D. Finite Mixture Models. Wiley, New York; 2000.
- 651 [27] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM
652 algorithm. Journal of the Royal Statistical Society Series B-Methodological. 1977;39(1):1-38.
- 653 [28] Schwarz G. Estimating the dimension of a model. Annals of Statistics. 1978;6:461-4.
- 654 [29] Diday E. Explanatory tools for machine learning in the symbolic data analysis framework. In:
655 Advances in Data Science: Symbolic, Complex and Network Data. vol. 4. ISTE Ltd and John
656 Wiley & Sons, Inc.; 2020. p. 3-30.
- 657 [30] Diday E. Introduction to the “s-concordance” and “s-discordance” of a Class with a Collection
658 of Classes. In: Analysis of Categorical Data from Historical Perspectives. Behaviormetrics:
659 Quantitative Approaches to Human Behaviour. vol. 17. Springer Singapore; 2023. Available
660 from: <https://doi.org/10.1007/978-981-99-5329-5-27>.
- 661 [31] IEA International Association for the Evaluation of Educational Achievement. PIRLS Progress
662 in International Reading Literacy Study; 2016. <https://timssandpirls.bc.edu/pirls2016>.
- 663 [32] Salton G, McGill MJ. Introduction to Modern Information Retrieval. New York: McGraw-Hill
664 Book Co.; 1983.
- 665 [33] Dobša J, Dalbelo-Bašić B. Comparison of information retrieval techniques: latent semantic
666 indexing and concept indexing. Journal of Information and Organizational Sciences. 2004;28(1-
667 2):1-15.