

Introduction to Symbolic Data Analysis

Uvod v simbolno analizo podatkov

Simona Korenjak-Černe

University of Ljubljana,
School of Economics and Business, and
Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia
`simona.cerne@ef.uni-lj.si`

NEW DEVELOPMENTS IN STATISTICS
SODOBNI STATISTIČNI PRISTOPI
Interdisciplinary Doctoral Programme in STATISTICS
13 May 2025

Outline

- 1 Symbolic Data Analysis (SDA)
 - Step 1: Data description
 - Step 2: Methods for analyzing symbolic data
 - SDA references
 - A brief insight into descriptive statistics in SDA
- 2 Clustering methodology for SDA
 - Symbolic object (SO) described with discrete distribution
 - `clamix` – clustering SO described with discrete distributions
 - Application: Clustering population pyramids
 - Application: Analyzing sex-age-cause specific mortality data
- 3 Symbolic concordance and discordance
 - Application on PIRLS data
- 4 Some related research fields
- 5 Invitation to the SDA Workshop 9 – 11 June 2025

Symbolic Data Analysis - SDA


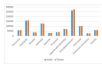

Symbolic Data Analysis (SDA) is an extension of the standard data analysis. SDA is a generalization of the traditional data analysis and provides complementary view on a data.

The need for such an extension stems from the idea of analysing aggregate data.

Originating from data science, SDA places great emphasis on data description. The data is represented in a **symbolic data table**, which preserves much more information about the individual data. Therefore, the traditional classical data matrix (where the value of a variable is a single point value) is replaced by **symbolic data table** to capture and represent more complex data in a more natural way.

An example of symbolic data table

More complex "values" of symbolic variables besides conventional single point values include also set of values, intervals, bar charts, histograms, functions, hierarchies, pictures, etc.

Units		Symbolic variables descriptions				
Country	Land size (km ²)	Flag	Population by regions	...	Births	...
⋮						
Slovenia	20273			...		...
⋮						

Besides the more intuitive data presentation, the aim is often also to analyse such data with appropriate tools (adjusted to data representation in order to preserve as much information as possible).

Initial idea for Symbolic Data Analysis

Symbolic Data Analysis was introduced in the eighties by Prof. Edwin Diday. It was originally based on the introduction of the **second level units** called **concepts**, which are usually a natural extension of aggregate descriptions of individuals. This type of data is called **symbolic** because it cannot be reduced to single numbers without loss of information (e.g., internal variation). They are inspired by Aristotle's works on logic (grouped in the collection known as *The Organon*), where he used the terminology of first-level objects called *individuals* and second-level objects.

Unlike classical data, symbolic data have **internal variation and structure** which should be taken into account when analysing such kind of data.



Diday, E. (1987): Introduction à l'approche symbolique en analyse des données. *Première Journées Symbolique-Numérique*, CEREMADE, Université Paris IX Dauphine, 21-56.



Diday, E. (1989): Introduction à l'Approche Symbolique en Analyse des Données. *Recherche opérationnelle/Operations Research* 23(2), 193–236. (in French)

Aggregated data as symbolic data

Aggregation

- is usually the first step to make large amount of data manageable;
- extracts (first) information from BIG DATA;
- protects privacy of individuals (persons, companies etc.);
- produces second level units of data (in SDA called classes).

Symbolic Data Analysis allows us to view aggregated data as new entities for further analysis at a higher level. Such **(second-level) unit** (called also **class**) preserves and allows us to include more information in further analysis (e.g., **internal variability**).

Main steps of SDA

Main steps of SDA are:

1. **Data description** (preferably presented in **the symbolic data table**)
Determine the new units of interest (concepts) and build the symbolic dataset. In practice, we have to solve the question how to formalize the concept and how to describe it to preserve as much information from data as possible. Often we add additional classical or symbolic valued variables of interest that relate to concepts in the symbolic database. Therefore, instead of using classical data matrix we present data in a **symbolic data table** that provides much more intuitive view on the data.
2. **Knowledge extraction from symbolic dataset**
by using descriptive statistics and other (more advanced) statistical and knowledge discovery tools adapted or newly developed for symbolic descriptions. Adaptations of traditional data analysis methods and development of new ones combine knowledge from different fields of science (mathematics, statistics, computer science, and others).

Data description

Symbolic object (SO) is a mathematical model of a concept.

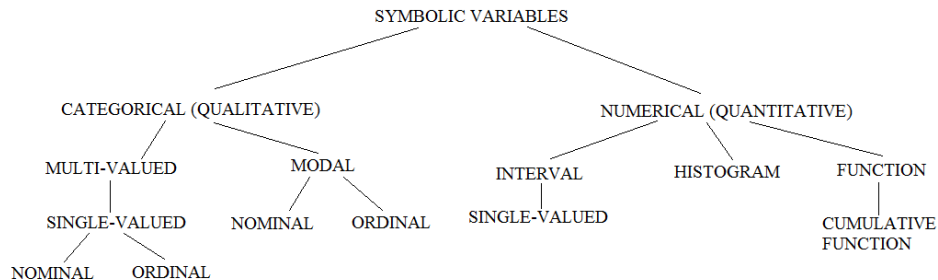
The symbolic observation very often represents the set of individuals corresponding to the description of the associated concept (e.g., when a new unit is obtained by aggregating individuals). In these cases, individuals represent the **first level units** and concepts represent the **second level units**, which are called **classes**.

BASIC NOTATION:

$Y_j(i) = x_{ij}$ a **classical value** or realization for the random variable $Y_j, j = 1, \dots, p$, on individual i (value x refers to the **individual**)

$Y_j(\omega_i) = \xi_{ij}$ a **symbolic value** or realization for the random variable $Y_j, j = 1, \dots, p$, on observation i (symbolic value ξ refers to the **class**)

Main types of symbolic variables considered in the SDA research so far



Source: Noirhomme-Fraiture, M., Brito, P. (2011): Far beyond the classical data models: symbolic data analysis. Fig. 1, p. 4.

Formal definitions of basic types of symbolic variables considered in the SDA research so far

Definition 1: A **multi-valued** symbolic variable Y is the one whose possible value takes one or more values from the list of values in its domain \mathcal{Y} . The complete list of possible values in \mathcal{Y} is finite or infinite, and values may be well-defined categorical (qualitative, nominal) or numerical (quantitative) values.

An example: Observed unit i : a person

selected variable: his/her car's type (Volvo, Renault, Fiat, Toyota,...)

Symbolic data: $\xi_i = \{\text{Volvo}, \text{Fiat}\}$ (if person i owns both types of cars)

Classical data: $x_i = \text{Volvo}$ or $x_i = \text{Fiat}$ (only one possible value)

Classical data - alternative: consider each possible value as a separated variable with values 'True' or 'False' (Volvo(i)=True, Renault(i)=False, Fiat(i)=True, ...)

Definitions ...

Definition 2: An **interval** symbolic variable Y is the one whose possible value is represented with an interval, i.e., $Y = \xi = [a, b] \subset \mathbb{R}$, with $a \leq b$, $a, b \in \mathbb{R}$. The interval can be closed or open at either end, i.e., $[a, b]$, (a, b) , $[a, b)$, $(a, b]$.

An example: Observed unit i : a person

selected variable: a range of credit card expenses in February in EUR

Symbolic data: $\xi_i = [42, 87]$

Classical data: $x_i = 64.5$ (only one possible value from the range, usually one of the mean values) – information of the range and internal variability is lost

Definitions ...

Definition 3: Let the random variable Y takes possible values $\{\eta_k, k = 1, 2, \dots\}$ over a domain \mathcal{Y} . Then, a particular outcome for an observation u is **modal valued** if it takes the form

$$Y(\omega_u) = \xi_u = \{(\eta_k, \pi_k); k = 1, \dots, s_u\},$$

where π_k is a non-negative measure (weight) associated with the outcome η_k from the domain $\mathcal{Y} = \{\eta_k, k = 1, 2, \dots\}$ and where s_u is the number of values actually taken from \mathcal{Y} .

In general, the η_k for modal-valued symbolic variable may be numerical (quantitative) or categorical (qualitative) in value and the domain \mathcal{Y} can be finite or infinite in size.

Most frequently, the weights π_k are frequencies or relative frequencies (probabilities); but they can also be capacities, or some other related forms.

SO examples: data described with modal nominal-valued variable

Examples:

1. Observed unit i : a flag
selected variable: color (red, blue, white, black,...)
Symbolic data: $\xi_i = \{\text{red}, 0.7; \text{white}, 0.3\}$ (if a flag i includes 70% of red color and 30% of white)
Classical data: $x_i = \text{red}$ or $x_i = \text{white}$ (only one possible value)
Classical data - alternative: consider each possible color as a separate variable with values between 0 and 1
2. Observed unit i : car dealer
selected variable: car type (Volvo, Toyota, Renault, ...)
Symbolic data: $\xi_i = \{\text{Renault}, 0.5; \text{Dacia}, 0.3; \text{Nissan}, 0.2\}$

Example for considering different units of observation: from individuals ...

descriptions of the individuals (first level units)

5. Observed individual i : a bird
 selected variables: Y_1 = species (ostrich, goose, penguin),
 Y_2 = flying (Yes/No),
 Y_3 = height (numerical)

Table 1: Birds: individual penguins, ostriches, and geese

Bird	Species	Flying	Height
1	Penguin	No	80
2	Goose	Yes	70
.	.	.	.
.	.	.	.
599	Penguin	No	80
600	Ostrich	No	125

Source: L. Billard, E. Diday: Symbolic Data Analysis, Table 2.18, p. 58

... to aggregated data

descriptions of the classes (second level units) extended with the additional variables

A concept (an entity of interest): species

$\omega_1 = \text{Penguin}$, $\omega_2 = \text{Geese}$, $\omega_3 = \text{Ostrich}$

selected (observed) variables: $Y_2 = \text{flying}$, $Y_3 = \text{height}$

additional (conceptual) variables: $Y_4 = \text{color}$, $Y_5 = \text{Migratory}$

Table 2: Concepts (species): penguin, geese, and ostrich

u	Species	Flying	Height	Color	Migratory
1	Penguin	No	[70,95]	{white, 0.5; black, 0.5}	Yes
2	Geese	Yes	[60,85]	{white, 0.3; black, 0.7}	Yes
3	Ostrich	No	[85,160]	{white, 0.1; black, 0.9}	No

Source: L. Billard, E. Diday: Symbolic Data Analysis, Table 2.19, p. 59

Definitions ...

Histograms are numerical (quantitative) modal-valued data, i.e., modal interval-valued data where η_k are nonoverlapping intervals and π_k are relative frequencies (probabilities).

Their realization has the form

$$Y = \{(I_k, p_k); k = 1, \dots, s\},$$

where I_k is the interval that can be open or closed at either end, s is the finite number of intervals for the domain of Y , and where p_k is the percentage of individuals in the particular subinterval I_k , $k = 1, 2, \dots, s$, $\sum_{k=1}^s p_k = 1$.

SO example: aggregated data

described with interval, multi-valued, histogram and modal ordinal-valued variables

Table: Data for healthcare centers

Healthcare center	Age of the patients Y_1	Number of emergency consults (on patient) Y_2	Pulse of the patients Y_3	Waiting time for consultations (min) Y_4	Education level of the patients Y_5
A	[25, 53]	{0, 1, 2}	[44, 86]	([0,15[(0),[15,30[(0.25), ([30,45[(0.5),[45,60[(0), ≥ 60 (0.25))	{9th grade, 1/2; Higher education, 1/2}
B	[33, 68]	{1, 4, 5, 10}	[54, 76]	([0,15[(0.25),[15,30[(0.25), ([30,45[(0.25),[45,60[(0.25), ≥ 60 (0))	{6th grade, 1/4; 9th grade, 1/4; 12th grade, 1/4; Higher education, 1/4}
C	[20, 75]	{0, 5, 7}	[70, 86]	([0,15[(0.33),[15,30[(0), ([30,45[(0.33),[45,60[(0), ≥ 60 (0.33))	{4th grade, 1/3; 9th grade, 1/3; 12th grade, 1/3}

Source: Noirhomme-Fraiture, M., Brito, P. (2011): Far beyond the classical data models: symbolic data analysis. Table 6, p. 6

Dependencies in the data: logical dependency rules

Rules are important for data coherence/integrity, putting condition(s) for underlying analysis, data cleaning etc.

Some examples:

- $Y_1 = \text{age}$, $Y_2 = \text{number of children}$;
aggregation of classical data $\xi = [5, 30] \times \{0, 1, 2, 3\}$;
the rule is needed, for example: $\{\text{If } Y_1 < 12 \text{ then } Y_2 = 0\}$
- $Y_1 = \text{annual income}$, $Y_2 = \text{tax}$;
If Y_1 is less than a certain threshold, $Y_2 = 0$
- $Y_1 = \text{have cancer}$, $Y_2 = \text{number of treatments}$;
If $Y_1 = \text{No}$ then $Y_2 = 0$

Some issues about symbolic data

Main question: which traditional terms can be extended with which properties?

If we allow more complex data descriptions where internal variability is preserved, several aspects need to be considered, for example,

- how can basic mathematical operations be performed (e.g. the sum of two intervals or two histograms)?
- how to obtain frequency distributions of intervals (with possibly different endpoints)?
- how to define mean value and standard deviation?
- how to define a linear combination of symbolic variables? For which types of symbolic variables can we define a vector space?

Methods for analyzing symbolic data

Tasks for knowledge extraction from symbolic datasets:

- Adaptations/generalizations of the classical statistical and knowledge discovery methods (for more complex data representations; taking into account the internal variation and structure(s));
- Developing new approaches for analyzing symbolic data.

A nice survey of the SDA methods can be found in the following presentation



Brito , P. (2012): Beyond summaries of individual data: Analyzing distributions. In *Symposium on Learning and Data Science*, Florence, Italy, 7 - 9 May 2012.

Methods for analyzing symbolic data ...

- Descriptive statistics
- Clustering:
 - based on dissimilarity measures
 - based on criterion (rules) of cluster description – conceptual clustering methods
- Multivariate statistical methods:
 - Regression (for interval-valued and histogram data)
 - Principal component analysis (for interval-valued and for modal multi-valued nominal data)
 - Discriminant analysis (for interval data)
 - Correspondence analysis
- Time-series analysis
- Visualization

SDA references

Main SDA books



Bock, H. H., Diday, E. (2000): *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data*. Springer, Heidelberg.



Billard, L., Diday, E. (2006): *Symbolic Data Analysis. Conceptual Statistics and Data Mining*. Wiley Series in Computational Statistics.



Diday, E., Noirhomme-Fraiture, M. (2008): *Symbolic Data Analysis and the SODAS Software*. John Wiley & Sons, Ltd.



Afonso, F., Diday, E., Toque, C. (2018): *Data science par analyse des données symboliques : Une nouvelle façon d'analyser les données classiques, complexes et massives à partir des classes*. Editions Technip. (in French)



Diday, E., Guan, R., Saporta, G., Wang, H. (2020): *Advances in Data Science: Symbolic, Complex and Network Data, Volume 4 - Big Data, Artificial Intelligence and Data Analysis*. ISTE and Wiley.



Billard, L., Diday, E. (2020): *Clustering Methodology for Symbolic Data*. Wiley Series in Computational Statistics.



Brito, P., Dias, S. (2022): *Analysis of Distributional Data*. Chapman and Hall/CRC.

SDA references

Some survey papers



Billard, L. (2006): Symbolic data analysis: what is it?. In: Rizzi A., Vichi M. (eds) *Compstat 2006 - Proceedings in Computational Statistics*. Physica-Verlag HD.



Noirhomme-Fraiture, M., Brito, P. (2011): Far beyond the classical data models: symbolic data analysis. In *Statistical Analysis and Data Mining* 4: 157–170.



Bruto, P. (2014): Symbolic Data Analysis: another look at the interaction of Data Mining and Statistics. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Volume 4, Issue 4: 281–295.



Diday, E. (2016): Thinking by classes in data science: the symbolic data analysis paradigm. In *Wiley Interdisciplinary Reviews: Computational Statistics*, Volume 8:172–205.

A brief insight into descriptive statistics in SDA

Comparison of symbolic and classical analysis - internal variability

Internal variation: Symbolic data has **internal variation and structure that should be considered when analyzing symbolic data**. Internal variability may be due to variability between lower-level units (for example, when the observed unit is obtained by aggregating individuals from the ground population), or it may be the result of observations of a single entity under different conditions (e.g., observing the time for a 100-meter dash of a particular athlete in different races or workouts).

If we omit the information about internal variability, very different units can have the same descriptions.

Illustrative example 1: $x = 5, \xi = [2, 8]$

Both samples have average 5. The classical sample variation for the single observation $x = 5$ is 0. The symbolic sample variation for the single observation $\xi = [2, 8]$, assuming uniform distribution within the interval, is 3.

Illustrative example 2: Intervals $[127, 133]$ and $[124, 136]$ have the same midpoint 130, but different internal variations, for example variance 3 and 12, respectively, when assuming uniform distributions within the intervals.

A brief insight into descriptive statistics in SDA

Comparison of symbolic and classical analysis ... proper unit selection

Statistical analysis on individuals and on classes are not the same, even though the same variable may be the object of the analysis. Special attention is needed to select the actual entity (unit) of interest.

Example: Relative frequencies of flying and non-flying birds and species:

600 birds, $\frac{2}{3}$ of them are flying and $\frac{1}{3}$ are non-flying.

3 species, $\frac{1}{3}$ of them are flying (Geese) and $\frac{2}{3}$ are non-flying (Ostrich and Penguin).

A brief insight into descriptive statistics in SDA

Comparison of symbolic and classical analysis ... frequencies

Special attention should be paid to the definition of frequencies of different interval values. For them, the frequencies are not necessarily integers, since we have to consider overlapping parts of the intervals. More detailed explanations with examples can be found in the chapters of the books Billard & Diday (2006) and Brito & Dias (2022).

A brief insight into descriptive statistics in SDA

Univariate statistics for interval and histogram variables with uniformly distributed values in the interval $[a_u, b_u]$ according to Bertrand and Goupil (2000) and Billard and Diday (2006)

INTERVAL VARIABLE Y

$$\text{Mean} \quad \bar{Y} = \frac{1}{n} \sum_{u \in E} \frac{b_u + a_u}{2}$$

$$\begin{aligned} \text{Variance} \quad S_Y^2 &= \frac{1}{3n} \sum_{u \in E} (b_u^2 + b_u a_u + a_u^2) - \frac{1}{4n^2} \left[\sum_{u \in E} (b_u + a_u) \right]^2 \\ &= \frac{1}{n} \sum_{u \in E} \left[\left(\frac{b_u + a_u}{2} \right)^2 + \frac{(b_u - a_u)^2}{12} \right] - \bar{Y}^2 \end{aligned}$$

HISTOGRAM VARIABLE

$$\text{Mean} \quad \frac{1}{2n} \sum_{u \in E} \left[\sum_{k=1}^{s_u} (b_{uk} + a_{uk}) \pi_{uk} \right] = \frac{1}{n} \sum_{u \in E} \left[\sum_{k=1}^{s_u} \frac{(b_{uk} + a_{uk})}{2} \pi_{uk} \right]$$

$$\text{Variance} \quad \frac{1}{3n} \sum_{u \in E} \left[\sum_{k=1}^{s_u} (b_{uk}^2 + b_{uk} a_{uk} + a_{uk}^2) \pi_{uk} \right] - \frac{1}{4n^2} \left[\sum_{u \in E} \left[\sum_{k=1}^{s_u} (b_{uk} + a_{uk}) \pi_{uk} \right] \right]^2$$

A brief insight into descriptive statistics in SDA

Bivariate statistics for interval variable with uniformly distributed values in the interval $[a_u, b_u]$ according to Billard and Diday (2006) and Billard (2008)

COVARIANCE $\text{Cov}(Y_j, Y_{j'})$

Billard and Diday (2003)
$$\frac{1}{4n} \sum_{u \in E} (a_{uj} + b_{uj}) (a_{uj'} + b_{uj'}) - \bar{Y}_j \bar{Y}_{j'}$$

Billard and Diday (2006)
$$\frac{1}{3n} \sum_{u \in E} G_{uj} G_{uj'} [Q_{uj} Q_{uj'}]^{\frac{1}{2}}, \text{ where}$$

$$Q_{uj} = (a_{uj} - \bar{Y}_j)^2 + (a_{uj} - \bar{Y}_j)(b_{uj} - \bar{Y}_j) + (b_{uj} - \bar{Y}_j)^2$$

$$G_{uj} = -1, \text{ if } \frac{a_{uj} + b_{uj}}{2} \leq \bar{Y}_j$$

$$G_{uj} = 1, \text{ if } \frac{a_{uj} + b_{uj}}{2} > \bar{Y}_j$$

Billard (2008)
$$\frac{1}{2} \sum_{u \in E} \frac{(b_{uj} - a_{uj})(b_{uj'} - a_{uj'})}{12}$$

$$+ \frac{1}{2} \sum_{u \in E} \left(\frac{a_{uj} + b_{uj}}{2} - \bar{Y}_j \right) \left(\frac{a_{uj'} + b_{uj'}}{2} - \bar{Y}_{j'} \right)$$

A brief insight into descriptive statistics in SDA

Univariate statistics for distributional (e.g., interval or histogram) symbolic variable according to Irpino and Verde (2015)

For the distributional symbolic variable (for example, interval-valued and histogram data) **Irpino and Verde (2015)** proposed definitions of the symbolic mean and symbolic variance based on the Mallows or L_2 Wasserstein (using naming from the authors) distance between two probability distributions ϕ_u and ϕ_v

$$d_W(\phi_u, \phi_v) = \sqrt{\int_0^1 [\Phi_u^{-1}(t) - \Phi_v^{-1}(t)]^2 dt},$$

where Φ_u^{-1} is a quantile function, i.e., the inverse of the *cdf* Φ_u (cumulative distribution function) that corresponds to the *pdf* ϕ_u (density function).

A brief insight into descriptive statistics in SDA

Squared L_2 Wasserstein distance



Irpino, A., Verde, R. (2015): Basic statistics for distributional symbolic variables: a new metric-based approach. In *Adv Data Anal Classif*, 9: 143–175.

R package `HistDAWas` - Histogram-Valued Data Analysis (Irpino, 2017)

If E is a set of n observed units described by the Y – distributional valued symbolic variable, such that the description of the u -th unit is $Y(u)$ with pdf ϕ_u having mean $\mu_u = E[Y(u)]$ and variance $\sigma_u^2 = E[(Y(u) - \mu_u)^2]$,

the correlation coefficient based on the two quantile functions Φ_u^{-1} and Φ_v^{-1} is defined as

$$\rho_{uv} = \frac{1}{\sigma_u \sigma_v} \left(\int_0^1 \Phi_u^{-1}(t) \Phi_v^{-1}(t) dt - \mu_u \mu_v \right).$$

Then, squared L_2 Wasserstein distance can be expressed as

$$d_W^2(\phi_u, \phi_v) = \underbrace{(\mu_u - \mu_v)^2}_{\text{location}} + \underbrace{(\sigma_u - \sigma_v)^2}_{\text{size}} + \underbrace{2\sigma_u \sigma_v (1 - \rho_{uv})}_{\text{shape}}$$

variability

A brief insight into descriptive statistics in SDA

Symbolic mean for distributional symbolic variable according to Irpino and Verde (2015)

Irpino and Verde defined symbolic mean as

$$M_W(Y) = \operatorname{argmin}_f \sum_{u \in E} d_W^2(\phi_u, f).$$

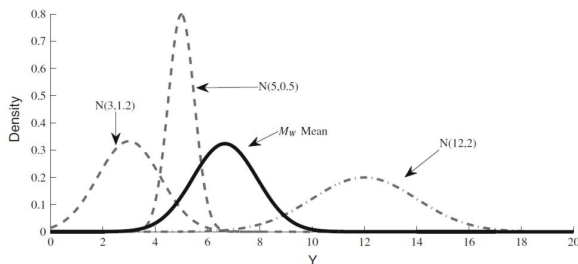


Fig. 1 The mean according to ℓ_2 Wasserstein distance

Source: Irpino, A., Verde, R. (2015): Basic statistics for distributional symbolic variables: a new metric-based approach. Fig.1

A brief insight into descriptive statistics in SDA

Symbolic variance for distributional symbolic variable according to Irpino and Verde (2015)

Symbolic variance based on L_2 Wasserstein distance is defined as

$$S_W^2(Y) = \underbrace{\left[\frac{1}{n} \sum_{u=1}^n \mu_u^2 - (\mu_{\bar{Y}})^2 \right]}_{SM_W^2(Y)} + \underbrace{\left[\frac{1}{n} \sum_{u=1}^n \sigma_u^2 - \frac{1}{n^2} \sum_{u=1}^n \sum_{v=1}^n \rho_{uv} \sigma_u \sigma_v \right]}_{SV_W^2(Y)}$$

where $SM_W^2(Y)$ represents the variability of means of the n distributions, and $SV_W^2(Y)$ is a measure of variance related to the differences of the internal variability of the n distributions.



Irpino, A., Verde, R. (2022): Descriptive Statistics for Numeric Distributional Data. In Brito, P., and Dias, S. (Eds.): *Analysis of Distributional Data*. CRC Press, Taylor & Frances Group.

Software

- SODAS: free, registration required
- SYR program: commercial software of the Symbad - Le Symbolic Data Lab
- R packages (in progress ...)
 - `HistDAWas` - Histogram-Valued Data Analysis (Irpino, 2017)
 - `RSDA` - An R package for symbolic data analysis (Murillo, Rodriguez, Villalobos, 2012)
 - `MAINT.Data` - Model and Analyze Interval Data (Duarte Silva, Brito; CRAN 2011)
 - `symbolicDA` - Analysis of symbolic data (Dudek, Pelka, 2012)
 - `R2S` - An R package to transform relational data into symbolic data (Murillo et al.)
 - `iRegression` - Some regression methods for interval-valued variables (de A. Lima Neto, de Souza Filho, Marinho, 2016)
 - `clustDDist` - Clustering Discrete Distributions,
`clamix` - Clustering Symbolic Objects (Batagelj, Kejžar; R-Forge 2010)

Clustering

Cluster analysis or **clustering** is the task of assigning a set of objects into groups called clusters such that the objects in the same cluster are *more similar* to each other than those in other clusters.

Different algorithms differ in their notion of what a cluster is and how to find them efficiently. Popular notions of clusters include groups with small distances between cluster members, dense regions of data space, intervals, or certain statistical distributions.

Clustering in SDA:



Billard, L., Diday, E. (2020): *Clustering Methodology for Symbolic Data*. Wiley Series in Computational Statistics.

Clustering methods in SDA for interval and histogram symbolic data - brief overview

Clustering methods for interval and histogram symbolic data:

- Irpino and Verde (2006): hierarchical and non-hierarchical clustering methods based on Wasserstein metric;
- De Carvalho and De Souza (2010): unsupervised pattern recognition models for mixed feature-type symbolic data;
- Brito and Ichino (2011): hierarchical clustering based on quantile representations;
- Brito and Chavent (2012): divisive algorithm for interval and histogram data.

clamix - Clustering methods for modal multi-valued nominal symbolic data

Clustering methods for modal multi-valued nominal symbolic data (with weights):

- Batagelj, Korenjak-Černe and Kejžar (2011+): hierarchical and non-hierarchical clustering methods based on discrete distributions (clamix).



Batagelj, V., Kejžar, N., Korenjak-Černe, S. (2015): Clustering of Modal Valued Symbolic Data. *ArXiv e-prints 1507.06683*, Jul 2015.



Kejžar, N., Korenjak-Černe, S., Batagelj, V. (2020): Clustering of modal-valued symbolic data. In *Adv Data Anal Classif*.



Batagelj, V., Korenjak-Černe, S., Kejžar, N. (2022): Clustering of Modal Valued Data. In: Brito, P., and Dias, S. (eds.) *Analysis of Distributional Data*, CRC Press.

Clustering software in SDA

Methods and their implementations in SODAS Software:

- 1 HIPYR Hierarchical and Pyramidal Clustering
- 2 DIV Divisive Clustering
- 3 SCLASS/VTREE Unsupervised Classification Tree
- 4 SCLUST Dynamic Clustering
- 5 DCLUST Clustering Algorithm based on Distance Tables
- 6 SYKSOM Kohonen Self-Organizing Map
- 7 CLINT Interpretation of Clusters

In R (in progress ...):

- 1 `symbolicDA`, `clusterSIM` (CRAN) (M. Walesiak, A. Dudek)
- 2 `clamix` - Clustering Symbolic Objects (V. Batagelj, N. Kejžar; R-Forge - work in progress)
- 3 `HistDAWas` - Histogram-Valued Data Analysis (A. Irpino)

Symbolic object (SO) described with discrete distribution

Clustering motivated by the real data:

- **Ego-centered networks:** include alter's information into ego's description (the values in the descriptions of different variables can be based on a different number and on different set of individuals - so called unpaired variables).
Example: TIMSS data set.
- **SOs represent demographic structures:** select **meaningful optimal cluster representative**, i.e., that represents the same demographic structure of the population of all units in the cluster.
Example: clustering population pyramids.
- **Patents' citation data set:** select **relative error measure** to reduce the influence of the largest component value.
Motivation for proposing alternative dissimilarities.

clamix – clustering SO described with discrete distributions

Classical clustering methods, adapted for SOs described with discrete distributions:

- *non-hierarchical leaders method* (a generalization of k-means method [Anderberg (1973), Hartigan (1975)], dynamic clouds [Diday (1979)])
- *agglomerative hierarchical clustering methods* (for example Ward's hierarchical clustering method [Ward (1963)])

clamix – program for clustering (very) large data sets of mixed units (units measured in different scales).



clamix – Clustering Symbolic Objects. R package. Authors: Batagelj, V. and Kejžar, N.

<https://r-forge.r-project.org/projects/clamix/>

clamix: data description based on discrete distribution

For the description based on distributions the domain of each variable

$V_i (i = 1, \dots, m)$ is partitioned into k_i subsets $\{V_{ij}, j = 1, \dots, k_i\}$.

The *set of units* \mathbf{U} consists of symbolic objects (SOs). An SO X is described with a list of descriptions of variables $V_i, i = 1, \dots, m$:

$$X = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m],$$

where m denotes the the number of variables and \mathbf{x}_i is a list of numerical values (usually frequencies or subtotals)

$$\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{ik_i}].$$

The same description (with the list of values for each symbolic variable) is used for cluster C .

clamix: data description based on discrete distribution ...

Such a representation of SO can be seen as based on *weighted modal* (or *histogram*, if $V_{ij}, i = 1, \dots, m, j = 1, \dots, k_i$, are intervals) type of symbolic variables in the following way:

let the sum of values of a variable V_i be denote with n_{x_i} :

$$n_{x_i} = \sum_{j=1}^{k_i} x_{ij}$$

and then calculate the corresponding empirical probability distribution with

$$\mathbf{p}_{x_i} = \frac{1}{n_{x_i}} \mathbf{x}_i = [p_{x_i1}, p_{x_i2}, \dots, p_{x_ik_i}] = [V_{i1}(p_{x_i1}), V_{i2}(p_{x_i2}), \dots, V_{ik_i}(p_{x_ik_i})].$$

Then, symbolic object X can be also described with a list of couples

$$X = [(n_{x_1}, \mathbf{p}_{x_1}), (n_{x_2}, \mathbf{p}_{x_2}), \dots, (n_{x_m}, \mathbf{p}_{x_m})].$$

clamix: data description based on discrete distribution - properties

- 1 it produces an **uniform description** for all types of variables;
- 2 we can **deal with variables** that are **based on a different number of original (individual) units**;
- 3 it requires a **fixed space** per variable. This is specially important if we are dealing with large data sets;
- 4 it is **compatible** with merging of disjoint clusters, i.e., knowing the descriptions of clusters C_1 and C_2 , $C_1 \cap C_2 = \emptyset$, we can easily calculate the empirical probability distribution of their union as a weighted sum;
- 5 the distributions in the description enable us to **reduce large data sets**;
- 6 it also **preserves more information** about the cluster than the usual value of an appropriate statistics – e.g., mean value used in the usual approach.

Ego-centered (personal) network as symbolic data

Ego-centered network

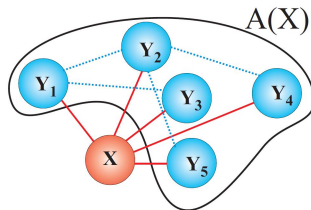


Figure: Ego-centered network

can be described with ego's and alters' variables (unpaired variables) as a symbolic object:

$$SO(X) = [X, A(X)]$$

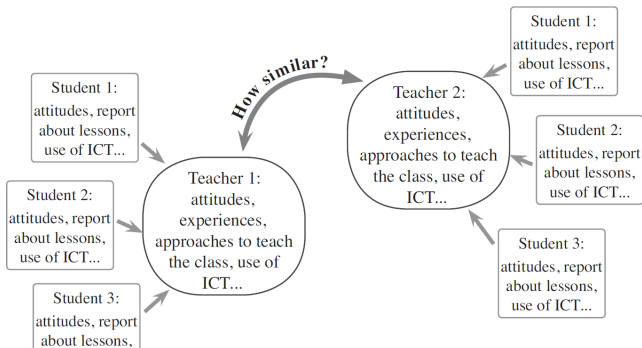
Example: TIMSS - data description

expert knowledge: Mrs. Barbara Japelj Pavešić, Slovenian TIMSS coordinator, The Educational Research Institute

TIMSS = Trends in International Mathematics and Science Study

<http://timss.bc.edu>, <http://www.pei.si/>

Units of the analysis were teachers, described by their variables: gender, age, education, their work in classes, pedagogical approaches used in the class, opinions about mathematics, classroom activities, use of IT and issues on homework.



Example: TIMSS - data description ...

Each teacher description also includes the distributions of student responses describing student attitudes toward mathematics, such as appreciation of mathematics, enjoyment of learning, confidence in mathematics, classroom activities such as use of IT, student engagement in mathematics learning, and their strengths in mathematics. These values were collected using the separate questionnaires for the students.

The symbolic object corresponding to teacher with id 4567 is

$$SO_{4567} = \left[\underset{\substack{\uparrow \\ T_1}}{(1, [0, 0, 0, 1])}, \quad \underset{\substack{\uparrow \\ T_2}}{(1, [0, 0, 0, 0, 1, 0])}, \quad \dots \quad \underset{\substack{\uparrow \\ S_1}}{(100, [0.47, 0.16, 0.37, 0, 0])}, \quad \underset{\substack{\uparrow \\ S_2}}{(100, [0, 0, 1, 0])}, \quad \dots \right]$$

where teacher variables (T_1, T_2, \dots) have only a singular value, but the alters' (students') variables S_1, S_2, \dots contain distributions of student's answers. Most of them are distributed over the following four subsets 1 = strongly agree, 2 = agree, 3 = disagree, and 4 = strongly disagree, that express how much they agree/disagree with the statement, that is considered as a student variable.

Optimization problem

Clustering as an optimization problem:

Find a clustering \mathbf{C}^* in a set of feasible clusterings Φ for which

$$P(\mathbf{C}^*) = \min_{\mathbf{C} \in \Phi} P(\mathbf{C}).$$

Model, assumed for leaders method: the criterion function is the sum of all cluster errors, the error of a cluster $p(C)$ is a sum of dissimilarities of its units from the cluster's *optimal representative – leader* T_C .

$$P(\mathbf{C}) = \sum_{C \in \mathbf{C}} p(C) \quad \text{where} \quad p(C) = \sum_{X \in C} d(X, T_C).$$

The set of feasible clusterings Φ is *a set of partitions* into k clusters of the finite set of units \mathbf{U} .

Optimization problem ...

We assume that the leader has the some structure of the description as SOs, i.e., it is represented with the nonegative vectors \mathbf{t}_i of the size k_i for each variable V_i – the representation space is $\mathcal{T} = (\mathbb{R}_0^+)^{k_1} \times (\mathbb{R}_0^+)^{k_2} \times \dots \times (\mathbb{R}_0^+)^{k_m}$.

For a given representative T and a cluster C we define the cluster error with respect to a representative T :

$$p(C, T) = \sum_{X \in C} d(X, T),$$

where d is a selected dissimilarity measure. The best representative T_C is called a *leader*

$$T_C = \arg \min_T p(C, T).$$

Then we define

$$p(C) = p(C, T_C) = \min_T \sum_{X \in C} d(X, T).$$

Optimization problem ...

We have to determine:

- **representations** of units, clusters, cluster representatives;
- **dissimilarity measures** between units, clusters, unit and cluster representatives.

A dissimilarity measure between SOs and T is defined as

$$d(X, T) = \sum_{i=1}^m \alpha_i d_i(\mathbf{x}_i, \mathbf{t}_i), \quad \alpha_i \geq 0, \quad \sum_{i=1}^m \alpha_i = 1,$$

where α_i are weights for variables (i.e., to be able to determine a more/less important variables) and

$$d_i(\mathbf{x}_i, \mathbf{t}_i) = \sum_{j=1}^{k_i} w_{x_{ij}} \delta(p_{x_{ij}}, t_{ij}), \quad w_{x_{ij}} \geq 0,$$

where $w_{x_{ij}}$ are weights for each variable's component.

Basic dissimilarities

Table clamix 1: The basic dissimilarities in the program

	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6
$\delta(x, t)$	$(p_x - t)^2$	$(\frac{p_x - t}{t})^2$	$\frac{(p_x - t)^2}{t}$	$(\frac{p_x - t}{p_x})^2$	$\frac{(p_x - t)^2}{p_x}$	$\frac{(p_x - t)^2}{p_x t}$

Related work (the application with the relative error measure):



Kejžar, N., Korenjak-Černe, S., Batagelj, V. (2011). Clustering of distributions: A case of patent citations, *Journal of Classification*, 28, 2, p. 156-183.

Non-hierarchical leaders method – the basic scheme of the algorithm

The leaders method as a variant of the dynamic clustering method:

determine an initial clustering \mathbf{C}_0 ; $\mathbf{C} := \mathbf{C}_0$

repeat

for each cluster C , $C \in \mathbf{C}$, determine its leader T_C ;

 assign each unit to the nearest new leader – producing a new clustering \mathbf{C}

until the leaders stabilize.

Determining new leaders

The new leader T_C of the cluster C is determined with

$$\begin{aligned} T_C &= \arg \min_T \sum_{X \in C} d(X, T) = \arg \min_T \sum_{X \in C} \sum_{i=1}^m \alpha_i d_i(X, T) = \\ &= \arg \min_T \sum_i \alpha_i \sum_{X \in C} d_i(\mathbf{x}_i, \mathbf{t}_i) = \left[\arg \min_{\mathbf{t}_i} \sum_{X \in C} d_i(\mathbf{x}_i, \mathbf{t}_i) \right]_{i=1}^m, \end{aligned}$$

where $\mathbf{t}_i = [t_{i1}, \dots, t_{ik_i}]$.

Because of the additivity of the model we can consider each variable separately and simplify the notation by omitting the index i .

Since in our general model the components are also independent, we can optimize component-wise and omit the index j :

$$t^* = \arg \min_{t \in \mathbb{R}} \sum_{X \in C} w_x \delta(p_x, t).$$

Optimal cluster representatives – leaders for different basic dissimilarities

Table clamix 2: The basic dissimilarities δ with the optimal cluster representative – leader t . C_t is the cluster with the leader t .

	$\delta(x, t)$	t	
δ_1	$(p_x - t)^2$	$\frac{P_t}{w_t}$	$w_t = \sum_{x \in C_t} w_x$
δ_2	$\left(\frac{p_x - t}{t}\right)^2$	$\frac{Q_t}{P_t}$	$P_t = \sum_{x \in C_t} w_x p_x$
δ_3	$\frac{(p_x - t)^2}{t}$	$\sqrt{\frac{Q_t}{w_t}}$	$Q_t = \sum_{x \in C_t} w_x p_x^2$
δ_4	$\left(\frac{p_x - t}{p_x}\right)^2$	$\frac{H_t}{G_t}$	$H_t = \sum_{x \in C_t} \frac{w_x}{p_x}$
δ_5	$\frac{(p_x - t)^2}{p_x}$	$\frac{w_t}{H_t}$	$G_t = \sum_{x \in C_t} \frac{w_x}{p_x^2}$
δ_6	$\frac{(p_x - t)^2}{p_x t}$	$\sqrt{\frac{P_t}{H_t}}$	

Application 1: TIMSS 1999, 2003

joint work with Slovenian TIMSS coordinator Barbara Japelj Pavešić from The Educational Research Institute

The goal was to:

- combine the dataset of teacher responses with the dataset of student responses;
- find groups of teachers with similar teaching approaches;
- check if there is a link between the obtained groups and student performance and/or some other variables.

Information about TIMSS data size for 2003:

Number of teachers (egos) = 6 552

Number of all students of the 8th grade included in the study (alters) = 131 000

Number of unpaired variables = 77 T + 24 S

Application 1: TIMSS 1999, 2003 ...

Clustering results

We have identified five main clusters. One of them contains units with predominantly missing values. Teachers in the other clusters differ in their use of computers and calculators in their lectures, in assigning and monitoring homework, and in testing their students' knowledge.

SYMBOLIC DESCRIPTIONS

The description of a leader for the cluster 1 by the teachers' variable T26:

T26 variable = "How often do you ask students to write equations during mathematics lessons?"

T26 name = btbmarr(MAT-HOW OFTEN AS-WRITE EQUATIONS)

Possible values: 'every or almost every lesson', 'about half the lessons', 'some lessons', 'never', 'missing', 'no sense'

$\text{card}(C_1) = 2\,050$, $t_{26}(C_1) = [0.052, 0.208, 0.689, 0.042, 0.011, 0]$

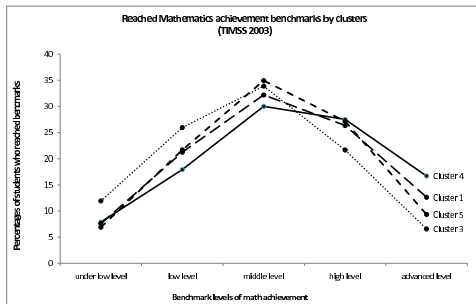
Some characteristics (variables with the highest percentage of teachers' answers) of the cluster 1

99.80%	btbmcoma(MAT-COMPUTER-AVAILABLE IN MATHS LESS) = YES
97.12%	btbmhmwo(MAT-HOMEWORK-DO YOU ASSIGN MATHS) = YES
97.07%	btbmtbtc(MAT-TEXTBOOK-USE FOR MATHS) = YES
85.17%	btbminta(MAT-COMPUTER-ACCESS TO INTERNET) = YES
84.05%	btbmased(MAT-HOW OFTEN ASK-INTERPRET DATA TABLES) = SOME LESSONS

Application 1: TIMSS 1999, 2003 ...

Links between the obtained clusters and students achievements

In the TIMSS study, students are assigned to different benchmark levels of mathematical knowledge. We also observed whether there are connections between the clusters obtained and student achievement (student achievement level was not considered in the clustering).



Benchmark levels of mathematics achievement reached by students
(cluster 2 with missing answers was omitted)

Additional details on the obtained results can be found in



Korenjak-Černe, S., Batagelj, V., Japelj-Pavešić, B. (2011): Clustering Large Data Sets Described with Discrete Distributions and its Application on TIMSS Data Set. In *Statistical Analysis and Data Mining, Special Issue on SDA*, Vol. 4, Issue 2, pp. 199–215.

Agglomerative hierarchical clustering method

The scheme of the standard agglomerative hierarchical clustering method:

each unit forms a cluster: $\mathbf{C}_n = \{\{X\}: X \in \mathbf{U}\}$;

they are at level 0: $h(\{X\}) = 0, X \in \mathbf{U}$;

for $k = n - 1$ **to** 1 **do**

 determine the closest pair of clusters

$(u, v) = \arg \min_{p, q: p \neq q} \{D(C_p, C_q): C_p, C_q \in \mathbf{C}_k\}$;

 join the closest pair of clusters $C_z = C_{(uv)} = C_u \cup C_v$

$\mathbf{C}_k = (\mathbf{C}_{k+1} \setminus \{C_u, C_v\}) \cup \{C_{(uv)}\}$;

$h(C_{(uv)}) = D(C_u, C_v)$

 determine the dissimilarities $D(C_{(uv)}, C_s), C_s \in \mathbf{C}_k$

endfor

\mathbf{C}_k is a partition of the finite set of units \mathbf{U} into k clusters. The level $h(C)$ of the cluster $C_{(uv)} = C_u \cup C_v$ is determined by the dissimilarity between the joint clusters C_u and C_v by $h(C_{(uv)}) = D(C_u, C_v)$.

Compatibility

The criterion function P is *compatible* with the dissimilarity D iff

$$(i) \quad P(\mathbf{C}) = \bigoplus_{C \in \mathbf{C}} p(C)$$

$$(ii) \quad p(C) = \min_{\emptyset \subset C' \subset C} (p(C') \oplus p(C \setminus C') \oplus D(C', C \setminus C'))$$

$$(iii) \quad p(\{X\}) = 0 \text{ for all } X \in \mathbf{U}$$

and $(\mathbb{R}_0^+, \oplus, 0, \leq)$ is an ordered Abelian monoid.

Table: Some compatible criterion functions and dissimilarities

METHOD	P	$p(C)$	$D(C_1, C_2)$
maximal (CL)	$\max p(C)$	$\max_{X, Y \in C} d(X, Y)$	$\max_{X \in C_1, Y \in C_2} d(X, Y)$
minimal (SL)	$\sum p(C)$	the value of the minimal spanning tree over C with edge values $d(X, Y)$	$\min_{X \in C_1, Y \in C_2} d(X, Y)$
Ward	$\sum p(C)$	$\sum_{X \in C} \ X - \bar{C}\ ^2$	$\frac{n_1 \cdot n_2}{n_1 + n_2} \ \bar{C}_1 - \bar{C}_2\ ^2$

$$n_i = \text{card}(C_i), \bar{C}_i = \frac{1}{n_i} \sum_{X \in C_i} X, \|\cdot\| \text{ is Euclidean norm.}$$

Compatible dissimilarities for different basic dissimilarities

Compatible dissimilarity in agglomerative hierarchical clustering:

$$D(C_u, C_v) = p(C_u \cup C_v) - p(C_u) - p(C_v)$$

Table clamix 3: The basic dissimilarities δ and the optimal cluster representative – leader t , the leader z of the merged clusters and dissimilarity D between merged clusters.

	$\delta(x, t)$	t	z	$D(C_u, C_v)$	
δ_1	$(p_x - t)^2$	$\frac{P_t}{w_t}$	$\frac{w_u u + w_v v}{w_u + w_v}$	$\frac{w_u \cdot w_v}{w_u + w_v} (u - v)^2$	$w_t = \sum_{x \in C_t} w_x$
δ_2	$(\frac{p_x - t}{t})^2$	$\frac{Q_t}{P_t}$	$\frac{u P_u + v P_v}{P_u + P_v}$	$\frac{P_u}{u} (\frac{u - z}{z})^2 + \frac{P_v}{v} (\frac{v - z}{z})^2$	$P_t = \sum_{x \in C_t} w_x p_x$
δ_3	$\frac{(p_x - t)^2}{t}$	$\sqrt{\frac{Q_t}{w_t}}$	$\sqrt{\frac{u^2 w_u + v^2 w_v}{w_u + w_v}}$	$w_u \frac{(u - z)^2}{z} + w_v \frac{(v - z)^2}{z}$	$Q_t = \sum_{x \in C_t} w_x p_x^2$
δ_4	$(\frac{p_x - t}{p_x})^2$	$\frac{H_t}{G_t}$	$\frac{H_u + H_v}{\frac{H_u}{u} + \frac{H_v}{v}}$	$G_u (u - z)^2 + G_v (v - z)^2$	$H_t = \sum_{x \in C_t} \frac{w_x}{p_x}$
δ_5	$\frac{(p_x - t)^2}{p_x}$	$\frac{w_t}{H_t}$	$\frac{w_u + w_v}{H_u + H_v}$	$w_u \frac{(u - z)^2}{u} + w_v \frac{(v - z)^2}{v}$	$G_t = \sum_{x \in C_t} \frac{w_x}{p_x^2}$
δ_6	$\frac{(p_x - t)^2}{p_x t}$	$\sqrt{\frac{P_t}{H_t}}$	$\sqrt{\frac{P_u + P_v}{\frac{P_u}{u^2} + \frac{P_v}{v^2}}}$	$\frac{P_u}{u} \frac{(u - z)^2}{uz} + \frac{P_v}{v} \frac{(v - z)^2}{vz}$	

Generalized Ward hierarchical method for clustering SOs described with discrete distributions

For the basic dissimilarity $\delta(x, t) = \delta_1(x, t) = (p_x - t)^2$ we get the generalization of the Ward dissimilarity

$$D(C_u, C_v) = \sum_{i=1}^m \alpha_i \sum_{j=1}^{k_i} \frac{w_{uij} \cdot w_{vij}}{w_{uij} + w_{vij}} (u_{ij} - v_{ij})^2$$

It also satisfies the Huygens theorem

$$I_T = I_B + I_W$$

Huygens theorem

Huygens theorem for δ_1 :

$$I_T = I_B + I_W$$

$$\text{Total inertia: } I_T = \sum_{X \in \mathbf{U}} d(X, T_U),$$

$$d(X, T_U) = \sum_i \alpha_i d(\mathbf{x}_i, \mathbf{t}_{Ui}) = \sum_i \alpha_i w_{xi} \|\mathbf{p}_{xi} - \mathbf{t}_{Ui}\|_2^2,$$

$$\mathbf{t}_{Ui} = \frac{1}{\sum_{X \in \mathbf{U}} w_{xi}} \sum_{X \in \mathbf{U}} w_{xi} \cdot \mathbf{p}_{xi}.$$

$$\text{Between inertia: } I_B = \sum_{C \in \mathbf{C}} d(T_C, T_U),$$

$$d(T_C, T_U) = \sum_i \alpha_i w_{Ci} \|\mathbf{t}_{Ci} - \mathbf{t}_{Ui}\|_2^2.$$

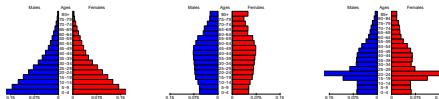
$$\text{Within inertia: } I_W = P(\mathbf{C}) = \sum_{C \in \mathbf{C}} p(C) = \sum_{C \in \mathbf{C}} \sum_{X \in C} d(X, T_C),$$

$$d(X, T_C) = \sum_i \alpha_i d(\mathbf{x}_i, \mathbf{t}_{Ci}) = \sum_i \alpha_i w_{xi} \|\mathbf{p}_{xi} - \mathbf{t}_{Ci}\|_2^2, \quad \mathbf{t}_{Ci} = \frac{1}{w_{Ci}} \sum_{X \in C} w_{xi} \cdot \mathbf{p}_{xi}, \text{ where } w_{Ci} = \sum_{X \in C} w_{xi}.$$

Clustering population pyramids

Motivation

Population pyramid



- is a popular representation of the age-sex distribution of a human population of a given region
- is influenced by
 - **population processes:** fertility, mortality, migration
 - **social and political policies:** birth control, wars, lifestyle
- pyramid **shape** reflects the characteristics of the observed time and region
- clusters of regions with similar shapes provide **additional insight** into this type of data

Population pyramid as SO

SO representation for `clamix`

X a region (world country, US county, municipality, subnational area) represented with population pyramid (age-sex distribution of population),

C_u a cluster of regions

Symbolic data description with two symbolic variables (one for each sex):

$$X = [(n_{xM}, \mathbf{p}_{xM}); (n_{xF}, \mathbf{p}_{xF})], \quad C_u = [(n_{uM}, \mathbf{p}_{uM}); (n_{uF}, \mathbf{p}_{uF})]$$

n_M - the number of men,

\mathbf{p}_M - relative frequencies of men over age groups,

n_F - the number of women,

\mathbf{p}_F - relative frequencies of women over age groups.

Population pyramid as SO ...

Data description - an example for Ljubljana

An example on Slovenian municipality Ljubljana:

The population of Ljubljana on 1st of July in 2011, split into 3 economic age groups 0-19, 20-64, 65+:

$$n_{LjM} = n_{men} = 134\,410, \quad n_{LjF} = n_{women} = 145\,488,$$

the corresponding frequency distributions:

$$[[25\,396, 90\,466, 18\,548], \quad [24\,204, 91\,899, 29\,385]]$$

and the corresponding data description with **symbolic object**

$$X_{Lj} = \left[(134\,410, [0.189; 0.673; 0.138]); \quad (145\,488, [0.166; 0.632; 0.202]) \right]$$

\uparrow
 n_{LjM}
 \uparrow
 p_{LjM}
}
 men

\uparrow
 n_{LjF}
 \uparrow
 p_{LjF}
}
 women

Clustering population pyramids with `clamix`

`clamix` – weighted clustering methods for discrete distributions (generalized k-means and Ward)

Adapted leaders method and **adapted Ward hierarchical method**, based on squared Euclidean distances **with** (possible) **weights** of sizes and/or of variables (Batagelj, V., Korenjak-Černe, S., Kejžar, N.).

For the cluster C_u it holds

$$n_{ui} = \sum_{X \in C_u} n_{xi} \quad \text{and} \quad p_{ui} = \frac{1}{n_{ui}} \sum_{X \in C_u} n_{xi} \cdot p_{xi}, \quad i = M, F.$$

The dissimilarity between clusters C_u and C_v is in this case rewritten into

$$D(C_u, C_v) = \frac{1}{2} \left(\frac{n_{uM} \cdot n_{vM}}{n_{uM} + n_{vM}} \|\mathbf{p}_{uM} - \mathbf{p}_{vM}\|^2 + \frac{n_{uF} \cdot n_{vF}}{n_{uF} + n_{vF}} \|\mathbf{p}_{uF} - \mathbf{p}_{vF}\|^2 \right).$$

Clustering population pyramids ...

clamix

The main characteristics of the method:

- 1 inclusion of weights for each variable (the number of males/females) in the clustering procedure
- 2 optimal cluster representative is the age-sex distribution (meaningful cluster representative)
- 3 age groups are considered as categories (not intervals)

We used the adapted hierarchical clustering method with the (weighted) squared Euclidean distance as dissimilarity for the applications on several open data sets

- *Slovenian municipalities* obtained from the National Statistical Office of the Republic of Slovenia.
- *population pyramids of the world countries*, obtained from the International Data Base (IDB).
- *US counties* from US Census 2000 and 2010 (with the additional variable ethnicity).
- *Brazilian municipalities* with IBGE - Brazilian Institute of Geography and Statistics data (use of age-sex and age-area (urban/rural) structures).
- *sub-national areas in Latin America and the Caribbean* with IPUMS dataset of census microdata from 1960 to 2011 (as part of the s-ALyC project).

Application 2: Clustering the world's countries with `clamix`

Using clustering to













- identify groups of countries with similar age-sex structures,
- identify groups of the world's countries with similar structural changes over time.

In order to achieve a relevant comparison, 215 of the world's countries for which data were available at all three time points (1996, 2001, 2006) were included in the analysis.

- We identified 4 main clusters for each of the observed years. Their shapes reflect the basic demographic stages of development quite well. The clusters were also examined in detail for partitions at lower levels.
- Observation over time showed that the shapes of the sex-age structures of the world's countries mostly changed from the more expansive shape to the stationary or even constrictive shape.
- Clustering based on temporal changes revealed five main clusters with similar temporal changes in the age-sex distributions of the population over the observed time.

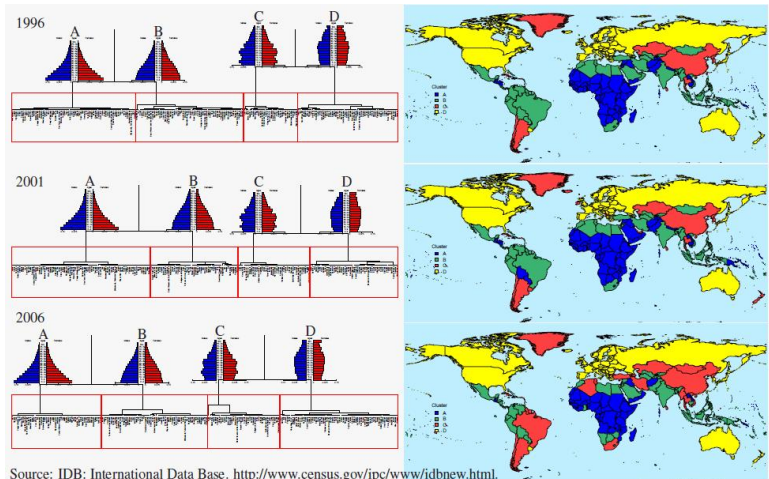
Application 2 ...

Four main clusters obtained from 215 of the world's countries for the years 1996, 2001, and 2006 with their population pyramids, number of countries in each cluster, and total population

Year	Description	A	B	C	D
1996	age-sex dist.				
	# countries	69	60	30	56
	population size	829,142,426	2,244,626,554	1,497,140,509	1,183,906,689
	% all population	14.4	39.0	26.0	20.6
2001	age-sex dist.				
	# countries	77	49	40	49
	population size	991,132,385	2,381,286,016	1,578,815,287	1,188,761,765
	% all population	16.1	38.8	25.7	19.4
2006	age-sex dist.				
	# countries	50	59	40	66
	population size	776,127,063	2,356,718,043	2,080,538,430	1,303,900,362
	% all population	11.9	36.2	31.9	20.0

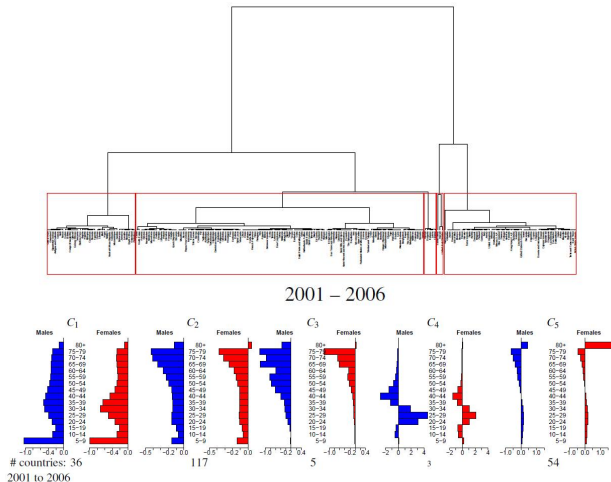
Application 2 ...

Four main clusters obtained from 215 of the world's countries for the years 1996, 2001, and 2006 presented in dendrograms and maps



Application 2 ...

Five main clusters obtained from 215 of the world's countries based on the age-cohort changes in 2001-2006



Application 2 ...

Highlights:

- unit representation with SO: saving complete information about age-sex distributions
- including weights for each variable (the number of males/females) in the clustering process
- age groups are considered as categories (not intervals)
- optimal cluster representative is the real age-sex distribution (meaningful cluster representative)
- detection of changes in each country considering common main clusters (with characteristic socio-demographic indicators)
- detection of clusters of the world's countries with similar changes in the age-sex structures over time

More detailed results are presented in



Korenjak-Černe, S., Kejžar, N., Batagelj, V. (2015). A weighted clustering of population pyramids for the world's countries, 1996, 2001, 2006. *Population Studies*, 69, 1, p. 105-120.

Clustering population pyramids using Wasserstein/Mallows L_2 Distance



Irpino, A., Lechevallier, Y., and Verde, R. (2006): Dynamic clustering of histograms using Wasserstein metric. In A. Rizzi, and M. Vichi (Ed): *COMPSTAT 2006*, 869–876. Berlin: Physica-Verlag.



Irpino, A. and Verde, R. (2006): A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In V. Batagelj, H.H., Bock, A. Ferligoj, and A. Žiberna (Ed): *Data Science and Classification*, 185–192. Berlin: Springer.



Košmelj, K., Billard, L. (2011): Clustering of Population Pyramids using Mallows' L_2 Distance. *Metodološki zvezki*, 2011, 8 (1), 1–15.

The main characteristics of the method:

- 1 histogram SO (use only relative age-sex structures, population size and number of males/females are not included in the clustering process)
- 2 include intervals of values for age groups
- 3 optimal cluster representative is barycentric histogram
- 4 cluster representative does not represent age-sex distribution

Application 3: Analyzing Sex-Age-Cause Specific Mortality in European Countries

Joint work with Aleša Lotrič Dolinar, Edwin Diday, Filipe Afonso and Jože Sambt.

Motivation: Knowing a country's position relative to others in terms of sex-age and sex-age-cause specific mortality is helpful in making decisions regarding health, demographic, and employment issues.

Data:

- Data 2014:

- 32 European countries (EU28, Liechtenstein, Norway, Serbia, Switzerland);
- original five-year age groups reduced into 7 age groups (0-14, 15-34, 35-54, 55-64, 65-74, 75-84, 85+) due to similar structures over causes of death;
- considering 3 main groups of causes of death (ICD-10) that represent over 70% of all deaths in European countries: circulatory, neoplasms, respiratory, all the others in a special category named others;
- mortality transformed on the standard population (size 100 000).

- Data 2015:

- 28 EU countries ;
- all 18 original five-year age groups;
- considering the same 3 main groups of causes of death: circulatory, neoplasms, respiratory, all the others in a special category named others;
- using 3-year averages for the number of deaths (to reduce the random effect);
- mortality transformed on the standard population (size 100 000).

Application 3 ...

Classical clustering of EU countries by causes of death

Classical approach:











































- Data 2015: 28 units described with 144 variables (2 sex x 18 age-groups x 4 categories-causes of death);
- Ward's and k-means clustering methods;
- Two main groups identified: Western and Eastern countries -> groups of countries are related to geographical location of countries;
- Mortality and cause of death are both discriminant, for both countries and clusters, for the different sex-age combinations.







Lotrič Dolinar, A., Sambt, J., Korenjak-Černe, S. (2019). Clustering EU Countries by Causes of Death. *Popul Res Policy Rev*, 38, 2, p. 157-172.

Application 3 ...

Part of the symbolic data table for sex-age-cause specific mortality in European countries in 2014

		Age groups						
Country	Sex	0-14	15-34	35-54	55-64	65-74	75-84	85 +
Austria (AT)	M	 2,7	 7,4	 34,6	 54,9	 95,3	 141,0	 117,0
	Ž	 2,2	 3,0	 18,3	 29,9	 59,2	 134,7	 237,2
Belgium (BE)	M	 3,2	 8,4	 36,2	 57,5	 94,5	 151,9	 122,8
	Ž	 2,3	 3,2	 22,5	 33,9	 60,1	 140,1	 227,4
Bulgaria (BG)	M	 6,8	 12,1	 72,7	 125,1	 176,5	 246,1	 155,6
	Ž	 4,9	 4,9	 33,4	 52,0	 98,1	 253,8	 311,0

 Neoplasms
  Circulatory
  Respiratory
  Other

Application 3 ...

Methodological issues for analyzing sex-age-cause specific mortality

Methodological issues:

- capture both dimensions for each sex-age combination:
 - structure over causes of death
 - and
 - mortality level;
- consideration of connections between classical variables (structures over causes of death for each sex-age combination).

Application 3 ...

Analyzing sex-age-cause specific mortality with SDA methods

SDA approaches:

- clamix – adapted Ward's and k-means for weighted clustering of relative structures;
- SYR program - two-stage-automatic data processing of symbolic analysis:
 - to achieve a suitable data representation that captures both dimensions equally (discretizing of mortality levels by each sex-age combination),
 - to group countries according to their mortality (using PCA for bar charts, adapted clustering and (automatic) detection of discriminating characteristics).



Lotrič Dolinar, A., Afonso, F., Korenjak-Černe, S., Diday, E. (2022). Complementary results on sex-, age-, and cause-specific mortality in EU countries obtained by SYR symbolic data analysis software. *Advances in methodology and statistics*, vol. 19, no. 1, p. 31-144.

S-concordance

Joint work with Edwin Diday

A "similarity" as a "concordance" in data analysis represents 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.

The s-concordance measures the similarity between an object and (with) a collection of objects.

S-concordance and s-discordance

The concordance that measures the similarity between an object and a collection of objects is based on classes described by symbolic data and falls within the framework of symbolic data analysis (SDA) (Diday, 2020), which is why it is called **s-concordance** (**symbolic concordance**).



Diday, E. (2023): Introduction to the "s-concordance" and "s-discordance" of a Class with a Collection of Classes. In *Analysis of Categorical Data from Historical Perspectives: Essays in Honour of Shizuhiko Nishisato* (Beh EJ, Lombardo R, Clavel JG, editors). Behaviormetrics: Quantitative Approaches to Human Behavior, vol 17, pp. 469 – 486, Springer Nature Singapore.

Note: This type of concordance differs from Kendal's concordance, which measures agreement between ordinal variables.

Symbolic concordance and discordance (intuitive)

- A class has **high concordance** with a given collection of classes for a category x if its frequency for x is frequent among the classes of that collection.
- A class has **high concordance** with a given collection of classes and with 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 the category x is also frequent.
- A class has **high discordance** with a given collection of classes for a category x if that category is frequent in that class and if, in addition, there are no (or at least not many) other classes in the given collection of classes for which the category x is also frequent.

In measuring symbolic concordance and discordance, two variabilities are considered: variability between individuals (with function f) and variability between classes (with function g).

Application on PIRLS data

Joint work with Edwin Diday and Barbara Japelj Pavešić

The Progress in International Reading Literacy Study (PIRLS) is an international assessment and research project measuring fourth-grade reading achievement and school and teacher practices related to instruction.

The reading achievement scale is obtained from several variables that **measure the quality of reading**. For more detailed information, see the reference provided.

We focus our study on survey data from 2016.



PIRLS Progress in International Reading Literacy Study (2016).

Examples of the use of s-concordance or s-discordance measures on PIRLS data

Examples of the use of s-concordance or s-discordance measures on PIRLS data

- Measuring how concordant is an observed country with all countries for the selected category based on g or additionally with weighting with f .
- Determination of the most deviating country/countries for the selected category, taking into account both the g and the f values.
- Determination of the most deviating school class(es) for the selected category, taking into account the g and f values.
- Comparison of a country's position compared to all countries in traditional paper-based reading versus digital online reading.

PIRLS 2016 International Benchmarks of Reading Achievement

PIRLS 2016 International Benchmarks of Reading Achievement
(Reading achievement scale: mean = 500, st.d. = 100):

Scale Score	International Benchmark
625	Advanced International Benchmark
550	High International Benchmark
475	Intermediate International Benchmark
400	Low International Benchmark

PIRLS 2016 International Benchmark Variable Codes:

Code	Description
1	below the Low International Benchmark
2	at or above the Low, but below the Intermediate International Benchmark
3	at or above the Intermediate, but below the High International Benchmark
4	at or above the High, but below the Advanced International Benchmark
5	at or above the Advanced International Benchmark

Source: PIRLS 2016 User Guide, Chapter 4, p. 63

Interpretation of f and g in our case

$f(x; c)$ (or shortened $f_c(x)$) is the proportion of individuals $w \in c$ for which the value of the categorical variable X is equal to x

In our illustrative example on PIRLS data, x is at least a high level of reading literacy and f stands for the proportion of students in a class c (country or school class) who achieve at least a high level of reading literacy.

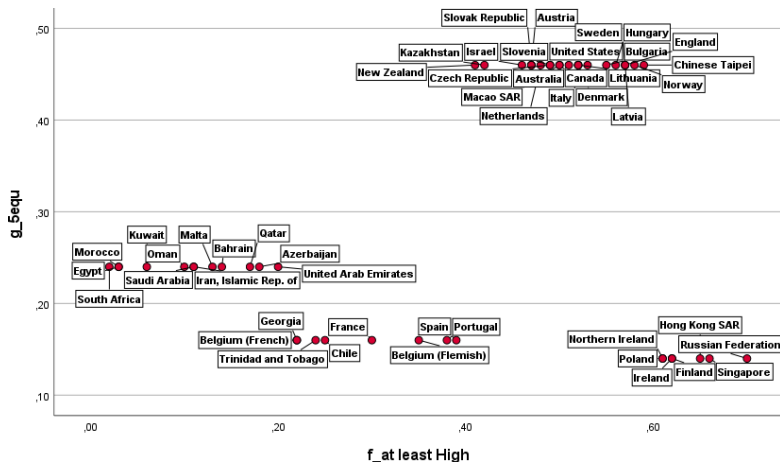
$g(c; P, x)$ (or shortened $g_x(c, P)$, where P is a collection of classes) is the proportion of classes c' from P for which the value $f(x; c')$ is equal or close to $f(x; c)$.

In our illustrative example, g stands for the proportion of classes from P in which the proportion of students with at least a high level of reading literacy is close to the proportion in class c .

By "close" we mean on the same subinterval of the interval $[0,1]$.

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

class = country, variable = international benchmark achievement in traditional (paper) reading, category = at least high level



s-discordance in text-mining

Joint work with Edwin Diday and Jasminka Dobša

Textual data: Comparison of TF-IDF (Term Frequency - Inverse Document Frequency) and s-discordance



Simona Korenjak-Černe, Jasminka Dobša, and Edwin Diday: An illustration of the use of the measures s-concordance and s-discordance in applications. Presented in the invited session *New Skills in Symbolic Data Analysis for Official Statistics* at the Conference on New Techniques and Technologies for Statistics NTTS2023, 6 - 10 March 2023, Brussels, Belgium.



Verde, R., Batagelj, V., Brito, P., Duarte Silva, P., Korenjak-Černe, S., Dobša, J., Diday, E. (2024). New skills in symbolic data analysis for official statistics. *Statistical Journal of the IAOS*, vol. 40, no. 3, p. 563-579.

Some related research fields

- CoDa - Compositional data analysis (Aitchison, 1986)
 - Non-negative data with constant sum (a quantitative description of the parts of some whole; the sample space of compositional data is a simplex)
 - density functions can be intuitively viewed as an infinite dimensional compositional data (Egozcue et al., 2006)
- FDA - Functional data analysis (Ramsay and Silverman, 2005, Menafooglio et al., 2018)



Aitchison, J. (1986): *The Statistical Analysis of Compositional Data*. Chapman and Hall.



Egozcue, J. J., Díaz-Barrero, J. L., Pawlowsky-Glahn, V. (2006): Hilbert space of probability density functions based on Aitchison geometry. In *Acta Mathematica Sinica* 22, 1175-1182.



Ramsay, J. O., Silverman, B. W. (2005): *Functional Data Analysis*. Springer.



Menafooglio, A., Secchi, P., Guadagnin, A. (2018, 2021): *Geostatistical analysis in Bayes spaces: probability densities and compositional data*. Wiley.

Summary

Symbolic Data Analysis (SDA) is an extension of standard data analysis. The traditional classical data matrix (where the value of a variable is a single point value) is replaced by a **symbolic data table** to capture and represent more complex data in a more natural way. The more complex "values" of symbolic variables include sets of values, intervals, bar charts, histograms, functions, hierarchies, images, etc.

Symbolic data have **internal variations and structures that should be considered** when analyzing this type of data. This requires adaptations and the development of new methods and techniques of data analysis. These adapted methods represent a **generalization of traditional data analysis methods** and provide a complementary view of the data. As such, they also enable the analysis of a large amount of data.

Invitation to the SDA Workshop 9 – 11 June 2025

Symbolic Data Analysis (SDA) Workshop 2025



WHEN: June 9 – 11, 2025

WHERE: Varaždin, Croatia, University of Zagreb, Faculty of Organization and Informatics (Foi), Varaždin

Workshop Highlights: Tutorial and Software Demonstrations: June 9, 2025

Online participation is available for the June 9th tutorial and software demonstrations, which is **free of charge**, only pre-registration is required.

The program for June 9:

9:30 - 11:00 Tutorial: Symbolic Data Analysis - Why, How, What for?
Paula Brito, University of Porto, Portugal

11:30 - 12:00 Software Presentation: The RSDA Package
Oldemar Rodriguez, University of Costa Rica

12:00 - 12:30 Software Presentation: R package MAINT.Data
Pedro Duarte Silva, Universidade Católica Portuguesa

12:30 - 13:00 Software Presentation: HistDAWass: an R package for the exploratory analysis of histogram data
Antonio Iripino, University of Campania Luigi Vanvitelli, Italy