Symbolic Data Analysis: Taking Variability in Data into Account

Paula Brito

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

ECI 2015 - Buenos Aires

Outline

Outline

The data

Classical data analysis :

Data is represented in a $n \times p$ matrix

each of n individuals (in row) takes one single value

for each of p variables (in column)

	Nb. children	Weight (Kg)	Gender	Education
Albert	2	52	M	2
Barbara	1	55	F	3
Charles	0	65	M	2
Deborah	3	60	F	1

The data

Symbolic variables :

to take into account **variability** inherent to the data

Variability occurs when we have

- Data about patients, but : analyse the healthcare centers - not the patients
- Data about flights, but : analyse the airports - not each individual flight
- Data about players, but : analyse the teams - not the individual players
- Data about people, but : analyse the parishes, the cities - not the individual citizens

Variable values are

sets, intervals

distributions on an underlying set of sub-intervals or categories

Micro-data $\longrightarrow$ **Macro-data**

The data

Example :
Data for three airports (e.g. arrival flights)

Airport	Number passengers	Delay arrival (minutes)	Aircraft
A	$[150, 200]$	$\{[0, 10[, 0.25; [10, 30[, 0.65; \geq 30, 0.10\}$	{Boeing 0.60 ; AirBus 0.40}
B	$[180, 300]$	$\{[0, 10[, 0.45; [10, 30[, 0.30; \geq 30, 0.25\}$	{Boeing 0.20 AirBus 0.80}
C	$[200, 400]$	$\{[0, 10[, 0.75; [10, 30[, 0.20; \geq 30, 0.05\}$	{Boeing 0.05 ; AirBus 0.95}

Sources of symbolic data

- Aggregation of micro-data: contemporary, temporal
- Description of abstract concepts

Sources of symbolic data: Aggregation of micro-data

Name	Ammount	Good	Card type
A	1000	drinks	Electron
A	4000	food	Visa
B	2000	food	Electron
A	15000	clothing	Mastercard
C	3000	food	Visa
B	2500	drinks	Electron
A	4000	food	Electron
C	7000	clothing	Mastercard
...	...	...	...

Temporal aggregation ↓

Name	Ammount	Good	Card type
A	[1000, 15000]	{drinks(1/4), food(1/2), clothing(1/4)}	{Electron, Visa, Mastercard}
B	[2000, 2500]	{drinks(1/2), food(1/2)}	{Electron}
C	[2000, 7000]	{food(1/2), clothing(1/2)}	{Visa, Mastercard}

Sources of symbolic data: Aggregation of micro-data

Communityname	State	perCapInc	pctPoverty	persPerOccupHous	pctKids2Par
Aberdeencity	SD	11939	12,2	2,35	76,25
Aberdeencity	WA	11816	18,3	2,34	64,05
Aberdeentown	MD	13041	10,66	2,61	60,79
Aberdeentownship	NJ	19544	3,18	2,86	79,31
Adacity	OK	10491	22,93	2,21	63,11
Adriancity	MI	11006	20,65	2,61	61,92
AgouraHillscity	CA	27539	3,53	3,08	86,65
Aikencity	SC	15619	15,69	2,48	64,51
Akroncity	OH	12015	20,48	2,42	55,76
Alabastercity	AL	13645	5,65	2,94	80,57
Alamedacity	CA	19833	6,81	2,36	70,29
...	...	...	...	...	...

Contemporary aggregation ↓

State	perCapInc	pctPoverty	persPerOccupHous	pctKids2Par
ALabama	[5820, 39610]	[2, 44]	[2, 3]	[30, 90]
ARkansas	[7399, 15325]	[4, 42]	[2, 3]	[45, 81]
AriZona	[6619, 62376]	[3, 43]	[2, 4]	[57, 90]
CAlifornia	[5935, 63302]	[1, 32]	[2, 5]	[47, 90]

Sources of symbolic data: Concept description

Description of the species "Dog" - not "my dog" !

Species	coat	vision range (m)	hearing frequency (Hz)	smell receptors (millions)
Dog	$\{single, double\}$	$[500, 900]$	$[40, 60000]$	$[125, 220]$

Outline

Symbolic Variable types

- Numerical (Quantitative) variables
  - Numerical single-valued variables
  - Numerical multi-valued variables
  - Interval variables
  - Histogram variables
- Categorical (Qualitative) variables :
  - Categorical single-valued variables
  - Categorical multi-valued variables
  - Categorical modal variables

Symbolic Variable types

$S = \{s_1, ..., s_n\}$: the set of n entities to be analyzed.
Let $Y_1, \ldots, Y_p$ be the variables, O_j the underlying domain of Y_j
B_j the observation space of $Y_j, j = 1, \ldots, p$

$$Y_j : S \longrightarrow B_j$$

- Y_j classical (numerical or categorical) single-valued variable : $B_j \equiv O_j$
- Y_j numerical or categorical multi-valued variable : $B_j = P(O_j)$
- Y_j interval variable : B_j set of intervals of O_j
- Y_j categorical modal or histogram variable : B_j set of distributions on O_j

Interval-Valued Variables

- $S = \{s_1, ..., s_n\}$: the set of n objects to be analyzed
- $Y_1, \ldots, Y_p$: the descriptive variables

Interval-valued variable :

$Y_j : S \rightarrow B : \ Y_j(s_i) = [l_{ij}, u_{ij}], l_{ij} \leq u_{ij}$
B : the set of intervals of an underlying set $O \subseteq \boldsymbol{R}$

I : $n \times p$ matrix - values of p interval variables on S

Each $s_i \in S$: represented by vector of intervals,
$I_i = (I_{i1}, ..., I_{ip}), i = 1, ..., n, \ I_{ij} = [l_{ij}, u_{ij}], j = 1, \ldots, p$

Interval data

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

Examples

Albert, Barbara and Caroline are characterized by the amount of time (in minutes) they need to go to work, which varies from day to day :

	Time
Albert	$[15, 20]$
Barbara	$[25, 30]$
Caroline	$[10, 20]$

Number of passengers in flights :

	Nb. Passengers
Airport A	$[150, 200]$
Airport B	$[180, 300]$
Airport C	$[200, 400]$

Native Interval Data

Temperatures and pluviosity measured in 283 meteorological stations in the USA:
temperature ranges in January and July, annual pluviosity range

Station	State	January Temperature	July Temperature	Annual Pluviosity
HUNTSVILLE	AL	[32.3, 52.8]	[69.7, 90.6]	[3.23, 6.10]
ANCHORAGE	AK	[9.3, 22.2]	[51.5, 65.3]	[0.52, 2.93]
NEW YORK (JFK)	NY	[24.7, 38.8]	[66.7, 82.9]	[2.70, 4.13]
...	...	...	...	...
SAN JUAN	PR	[70.8, 82.4]	[76.9, 87.4]	[2.14, 6.17]

Distribution-Valued Data

Keeping more information (requires more data at the micro level)
Example : Data for three airports

Airport	Delay arrival (minutes)	Aircraft
A	$\{[0, 10[\ , 0.25; [10, 30[\ , 0.65; \geq 30, 0.10\}$	{Boeing 0.60 ; AirBus 0.40}
B	$\{[0, 10[\ , 0.45; [10, 30[\ , 0.30; \geq 30, 0.25\}$	{Boeing 0.20 AirBus 0.80}
C	$\{[0, 10[\ , 0.75; [10, 30[\ , 0.20; \geq 30, 0.05\}$	{Boeing 0.05 ; AirBus 0.95}

Histogram-valued variables

Histogram-valued variable : $Y_j : S \rightarrow B_j$

B_j : set of probability or frequency distributions in a set of sub-intervals $\left\{I_{ij1}, ..., I_{ijk_j}\right\}$

$$Y_j(s_i) = (I_{ij1}, p_{ij1}; \ldots; I_{ijk_j}, p_{ijk_j})$$

$p_{ij\ell}$: probability or frequency associated to $I_{ij\ell} = [\underline{I}_{ij\ell}, \overline{I}_{ij\ell}[$
$p_{ij1} + \ldots + p_{ijk_j} = 1$

$Y_j(s_i)$ may be represented by the histogram :

$$H_{Y_j(s_i)} = ([\underline{I}_{ij1}, \overline{I}_{ij1}[, p_{ij1}; \ldots; [\underline{I}_{ijk_j}, \overline{I}_{ijk_j}], p_{ijk_j})$$

Histogram-valued variables

- Assumption : within each sub-interval $[\underline{l}_{ij\ell}, \overline{l}_{ij\ell}[$ the values of variable Y_j for observation s_i, are uniformly distributed
- For each variable Y the number and length of sub-intervals in $Y_j(s_i)$, $i = 1, \ldots, n$ may be different
- Interval-valued variables : particular case of histogram-valued variables: $Y_j(s_i) = [l_{ij}, u_{ij}] \rightarrow H_{Y_j(s_i)} = ([l_{ij}, u_{ij}], 1)$

Histogram-valued variables

$Y_j(s_i)$ can, alternatively, be represented by the inverse cumulative distribution function - quantile function

$$q_{ij}(t) = \begin{cases} \underline{l}_{ij1} + \frac{t}{w_{ij1}} r_{ij1} \ \textit{if} \ 0 \leq t < w_{ij1} \\ \underline{l}_{ij2} + \frac{t - w_{ij1}}{w_{ij2} - w_{i1}} r_{ij2} \ \textit{if} \ w_{ij1} \leq t < w_{ij2} \\ \vdots \\ \underline{l}_{ijK} + \frac{t - w_{ijK-1}}{1 - w_{ijK-1}} r_{ijK} \ \textit{if} \ w_{ij\ K-1} \leq t \leq 1 \end{cases}$$

where $w_{ijh} = \sum_{\ell=1}^{h} p_{ij\ell}, h = 1, \ldots, K; r_{ij\ell} = \overline{l}_{ij\ell} - \underline{l}_{ij\ell}$ for $\ell = \{1, \ldots, K\}$.

Histogram-valued variables: Example

Studying the performance of companies - delay time in arrival for some flights:

Company	Delay in arrival (minutes)
A	$5, 10, 15, 17, 20, 20, 25, 30, 30, 32, 35, 40, 40, 45, 50, 50$
B	$5, 8, 10, 12, 15, 20, 25, 25, 30, 32, 35, 35, 45, 52, 55, 60$

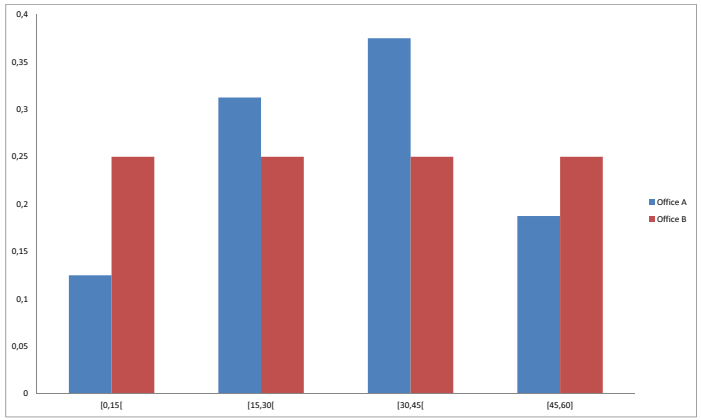
Average delay time : 29.0 minutes for both companies

Description in terms of histograms :

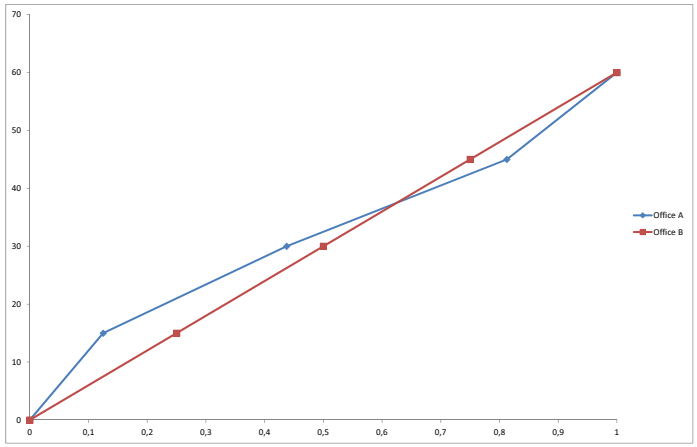
Company	Delay in arrival (minutes)
A	$\{[0, 15[, 0.125; [15, 30[, 0.3125; [30, 45[, 0.375; [45, 60], 0.1875\}$
B	$\{[0, 15[, 0.25; [15, 30[, 0.25; [30, 45[, 0.25; [45, 60], 0.25\}$

Histogram-valued variables: Example

Histograms :



Quantile functions :



Categorical modal variables

Categorical modal variable Y_j with a finite domain $O_j = \{c_{j1}, \ldots, c_{jk_j}\}$:

multi-state variable for each element: a category set

for each category $c_{j\ell}$, a weight, frequency or probability indicating how frequent or likely that category is for this element.

If the weight is a frequency :
proportion of individuals of the underlying microdata set characterized by this category

$$Y_j(s_i) = \{c_{j1}(p_{ij1}), \ldots, c_{jk_j}(p_{ijk_j})\},\ p_{i1} + \ldots + p_{ijk_j} = 1.$$

Categorical modal variables: Example

Stydying and comparing university departments, data about the faculty of each department.

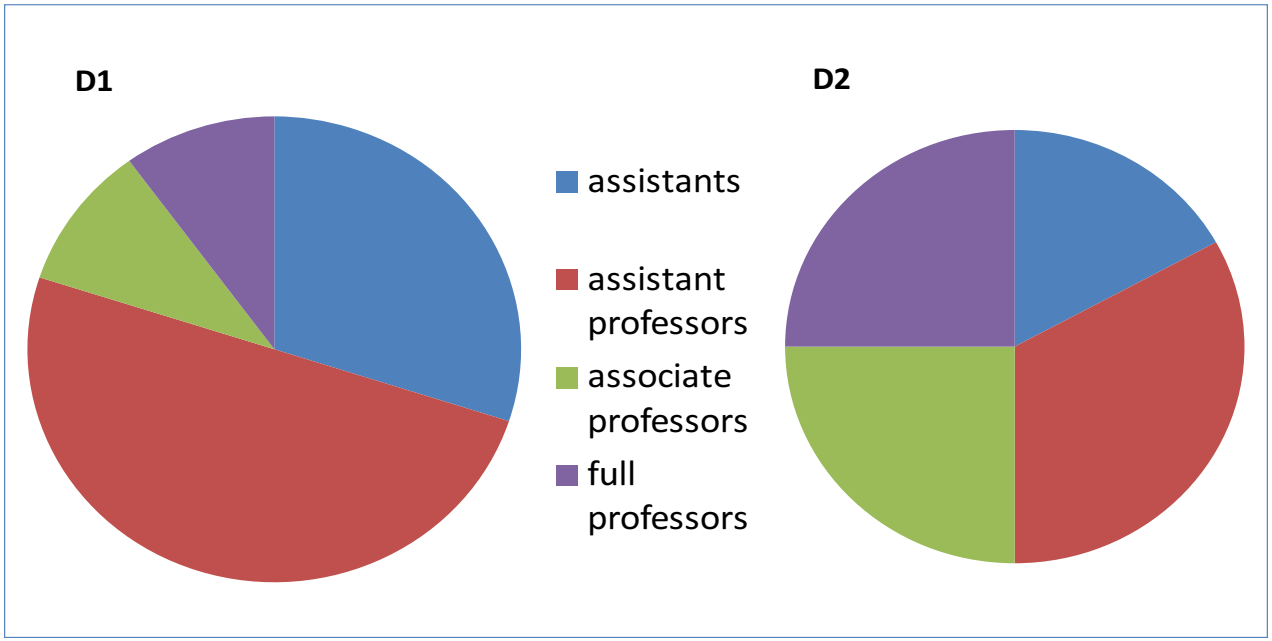
D_1 : 3 assistants, 5 assistant professors, 1 associate professor and one full professor
D_2 : 2 assistants, 4 assistant professors, 3 associate professors and 3 full professors

Taking modes:
departments mainly composed by assistant professors.

Categorical modal variable, $c_1 =$"assistant", $c_2 =$ "assistant professor", $c_3 =$"associate professor", $c_4 =$"full professor":

$$Y(D_1) = \{c_1(0.3), c_2(0.5), c_3(0.1), cm_4(0.1)\}$$
$$Y(D_2) = \{c_1(0.17), c_2(0.33), c_3(0.25), c_4(0.25)\}$$

Categorical modal variables: Example



Symbolic Data Array

A symbolic data array is represented in an $n \times p$ matrix, where

- Each row corresponds to an entity $s_i, i = 1, \ldots, n$
- Each column corresponds to a symbolic variable, $Y_j, j = 1, \ldots, p$
- To each row corresponds a **description**: $d_i = (Y_1(s_i), \ldots, Y_p(s_i))$

In the example above,
$([1000, 15000], \{\text{drinks}(1/4), \text{food}(1/2), \text{clothing}(1/4)\}, \{\text{Electron, Visa, Mastercard}\})$
is the description of individual A.

Symbolic Data Array

Data for arrival flights at three airports for each flight the number of passengers, the delay time (in minutes), and the distance category (say, from 1-domestic flight to 5-very long distance intercontinental flight) have been recorded

Data for each flight has then been aggregated by airport:

Airport	Number passengers	Delay time on arrival (minutes)	Distance category
A	$[150, 200]$	$\{[0, 10[, 0.25; [10, 30[, 0.65; [30, 60], 0.10\}$	{ 1 (0.40); 2 (0.40); 3 (0.2)}
B	$[180, 300]$	$\{[0, 10[, 0.45; [10, 30[, 0.30; [30, 60], 0.25\}$	{ 1 (0.10); 2 (0.30); 3 (0.30); 4 (0.20); 5 (0.10)}
C	$[200, 400]$	$\{[0, 10[, 0.75; [10, 30[, 0.20; [30, 60], 0.05\}$	{ 1 (0.05); 2 (0.10); 3 (0.15); 4 (0.40); 5 (0.30)}

Outline

Quantile representation

- **Objective**: Obtain a common representation model for different variable types
- Allowing to apply multivariate methods to the full (originally) mixed data array
- Discrete approach : For each observation $Y_j(s_i)$ - use the m-quantiles of the underlying distribution of the observed data values $(min; Q_1; \ldots; Q_{m-1}; max)$ (Ichino, 2008)
- When quartiles are chosen $(m = 4)$: representation for each variable is defined by the 5-uple $(min; Q_1; Q_2; Q_3; max)$
- Determination of quantiles for each variable type
- Continuous approach: determine quantile functions for each observation $Y_j(s_i)$

Quantile representation

- Interval-valued variables
  - Uniform or other distribution assumed in each interval
- Histogram-valued variables
  - Quantiles obtained by interpolation
  - Uniform distribution assumed in each class (bid)
- Categorical modal (or multi-valued) variables
  - Y_j categorical multi-valued variable taking possible k categories $c_\ell, \ell = 1, 2, \ldots, k$
  - Rank the categories $c_1, c_2, \ldots, c_k$ according to **a)** given order OR **b)** e.g., frequency values p_ℓ
  - Define a uniform cumulative distribution function for each object $s_i \in S$ based on the ranking, assuming continuity
  - Then find the $m - 1$ quantile values

Quantile representation: Example

Oils and Fats data

Oil or Fat	Specific gravity (g/cm^3)	Freezing point ($^\circ C$)	Iodine value	Saponif. value	Major acids
Linseed	$[0.930, 0.935]$	$[-27, -18]$	$[170, 204]$	$[118, 196]$	L, Ln, O, P, M
Perilla	$[0.930, 0.937]$	$[-5, -4]$	$[192, 208]$	$[188, 197]$	L, Ln, O, P, S
Cotton	$[0.916, 0.918]$	$[-6, -1]$	$[99, 113]$	$[189, 198]$	L, O, P, M,S
Sesame	$[0.920, 0.926]$	$[-6, -4]$	$[104, 116]$	$[187, 193]$	L, O, P, S, A
Camelia	$[0.916, 0.917]$	$[-21, -15]$	$[80, 82]$	$[189, 193]$	L, O
Olive	$[0.914, 0.919]$	$[0, 6]$	$[79, 90]$	$[187, 196]$	L, O, P, S
Beef	$[0.860, 0.870]$	$[30, 38]$	$[40, 48]$	$[190, 199]$	O, P, M, C,S
Hog	$[0.858, 0.864]$	$[22, 32]$	$[53, 77]$	$[190, 202]$	L, O, P, M, S, Lu

Quantile representation: Example

Linseed	Spec. Grav.	Freezing P.	Iodine	Saponific.	M. Acids
Min	0.93000	-27	170	118	4
Q1	0.93125	-24.75	178.5	137.5	5.25
Q2	0.93250	-22.5	187	157	7.5
Q3	0.93375	-20.25	195.5	176.5	8.75
Max	0.93500	-18	204	196	10

Outline

Interval data : Survey data application

Gender, Age, Level of Education, Job Category,
Income and debt variables - Household Income (HI), Debt to Income Ratio ($\times$ 100) (DIR), Credit Card Debt (in thousands) (CCD), Other Debts (OD)

5000 observations:

Gender	Age	Education	Job	HI	DIR	CCD	OD
Male	22	High school degree	Services	40	10	3	2
Male	45	College degree	Sales and Office	100	15	8	7
Female	30	Some college	Managerial and Professional	50	20	2	1

Individual observations aggregated on the basis of
Gender , Age Category , Level of Education and Job Category

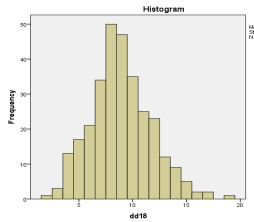
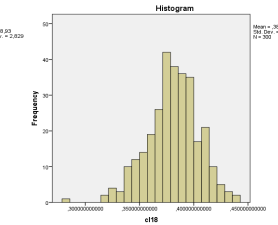
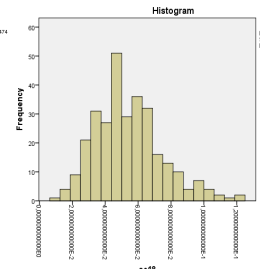
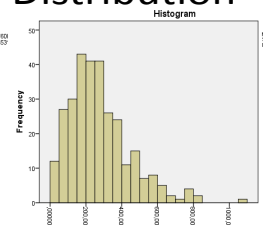
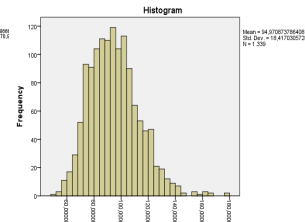
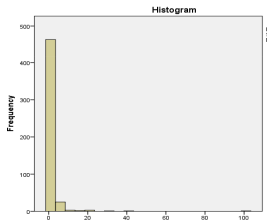
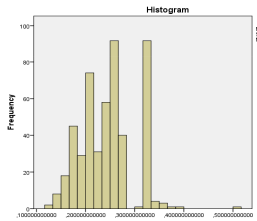
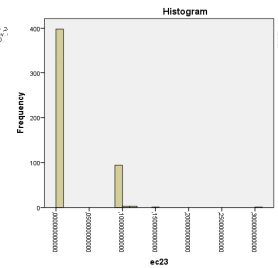
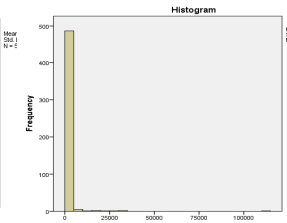
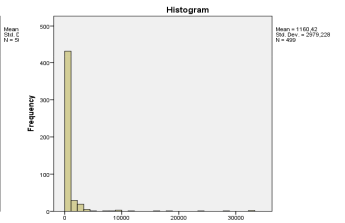
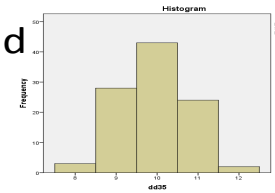
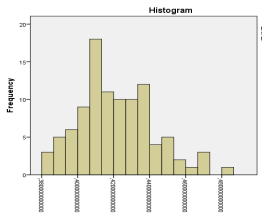
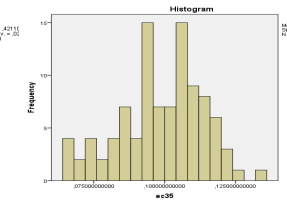
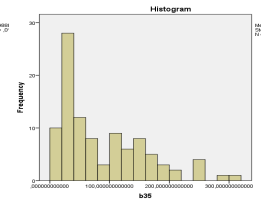
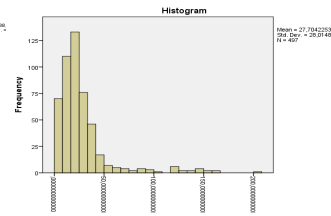
Interval data: Survey data application

Group	HI	DIR	CCD	OD
Male, 18-24 High school degree, Service	[15, 61]	[0.1, 23.4]	[0.0, 6.57]	[0.02, 7.71]
Male, 35-49, College degree, Sales and Office	[19, 190]	[1.4, 20.4]	[0.04, 16.6]	[0.12, 15.39]
Female, 25-34, Some college Managerial and Professional	[17, 100]	[0.8, 31.7]	[0.05, 6.57]	[0.09, 7.65]

Social Networks application (Giordano and Brito (2012))

- Describe each network by a vector of distributions of network indices
- One network $\rightarrow$ histogram-valued data vector
- Allows analysing networks represented as symbolic data (clustering,...)

Social network data as histogram data

	Degree Distrib.	Closeness Distribution	Eigenvector Distribution	Betweenness Distribution	Edge Betw. Distribution
Random Graph N=300 p=0,03					
Preferential Attachement N=500 λ=1,25					
Small World N=100 π=0,05					

Applications

In general : when it is wished to analyse data at a higher level (groups), rather than at individual level

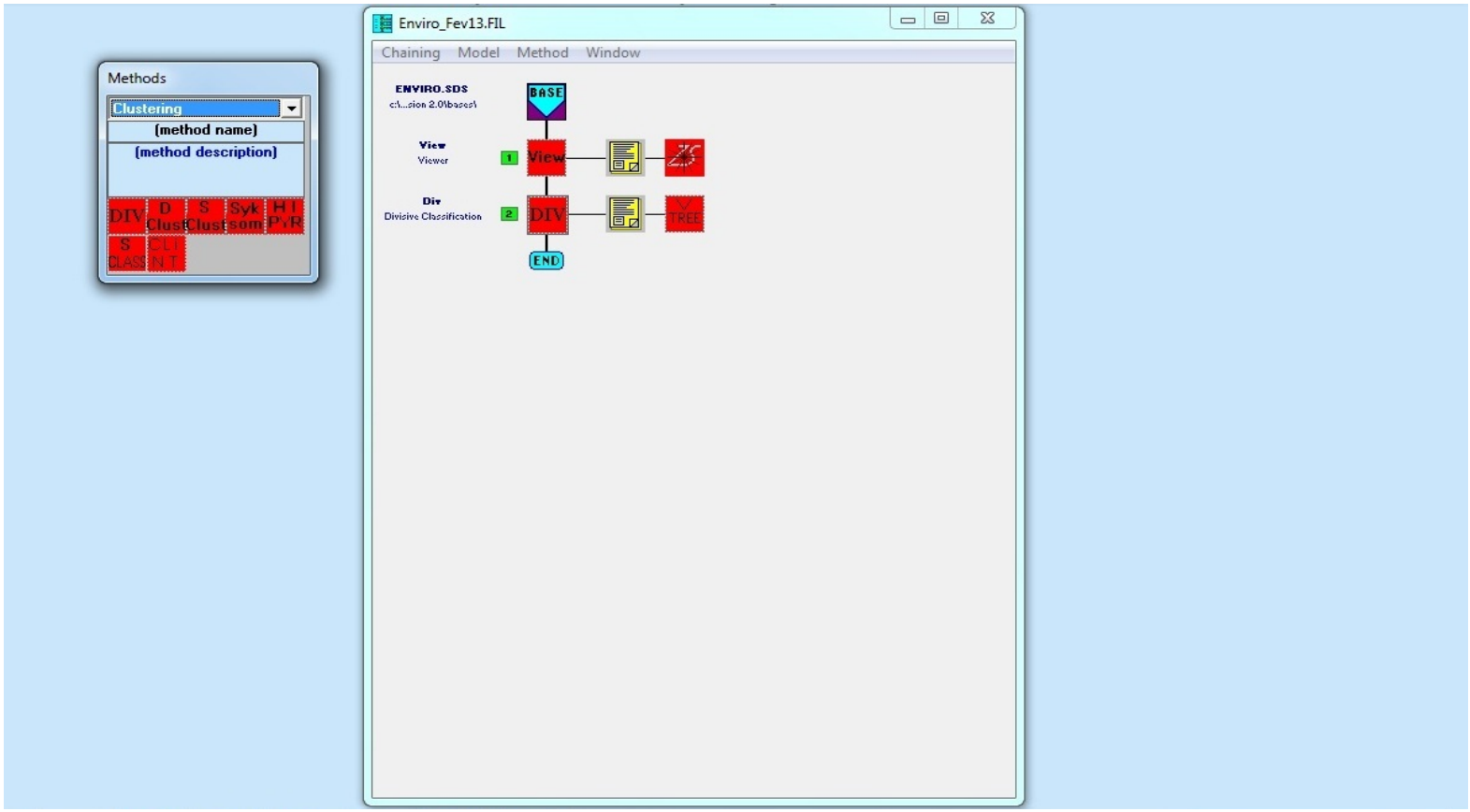
- Official data: confidentiality issues $\rightarrow$ aggregation
- Survey data
- Big databases, e.g., purchases per client, phone calls per person, prescriptions per patient or per doctor
- Analysis of abstract concepts as such
- ...

Outline

SODAS

- Build a symbolic data array from an Access database: DB2SO
  - Data is aggregated, based on given properties - e.g. : flights aggregated by companies
  - Numerical data: interval or histogram-valued data
  - Categorical data: set-valued or categorical-modal data
- Univariate / Bivariate analysis
- Multivariate Data Anaysis: clustering, factorial analysis, discriminant analysis, regression, etc.

SODAS


Methods
Clustering
(method name)
(method description)
DIV
D Clust
S Clust
Syk som
H I PyR
S CLAS
CLI N T
Enviro_Fev13.FIL
Chaining
Model
Method
Window
ENVIRO.SDS
BASE
View
Viewer
1
View
Div
Divisive Classification
2
DIV
TREE
END

SODAS

- Methods "window"
- Chaining
  - Sodas database
  - Methods application
    - listings
    - graphics
  - END

SODAS: data files

SODAS data files :

- xxx.sds
- xxx.XML

Data visualization:

- Data array
- Zoom star : 2D ; 3D

SODAS: Data example

Survey on Environmental Attitude EAS 2000

- Survey occasionally conducted earlier 1983, 1989 and 1994 by Statistics Finland
- The population is Finish resident aged from 15 to 74 in December 2000
- The sample contains 2500 persons and the collect is organized by interview and missing data are corrected by logistic regression
- The survey contains 35 questions, only 17 questions are selected :
- 4 categorical - urbanicity, incomelevel, education, regiondevelopment - 13 numerical

SODAS: Data example

Survey on Environmental Attitude EAS 2000

- The set of symbolic objects is build by Cartesian product between 3 categorical variables :
- Gender (2 categories)
- Age (4 categories) [15-24 , 25-44, 45-64, 65-74])
- Urban (2 categories)

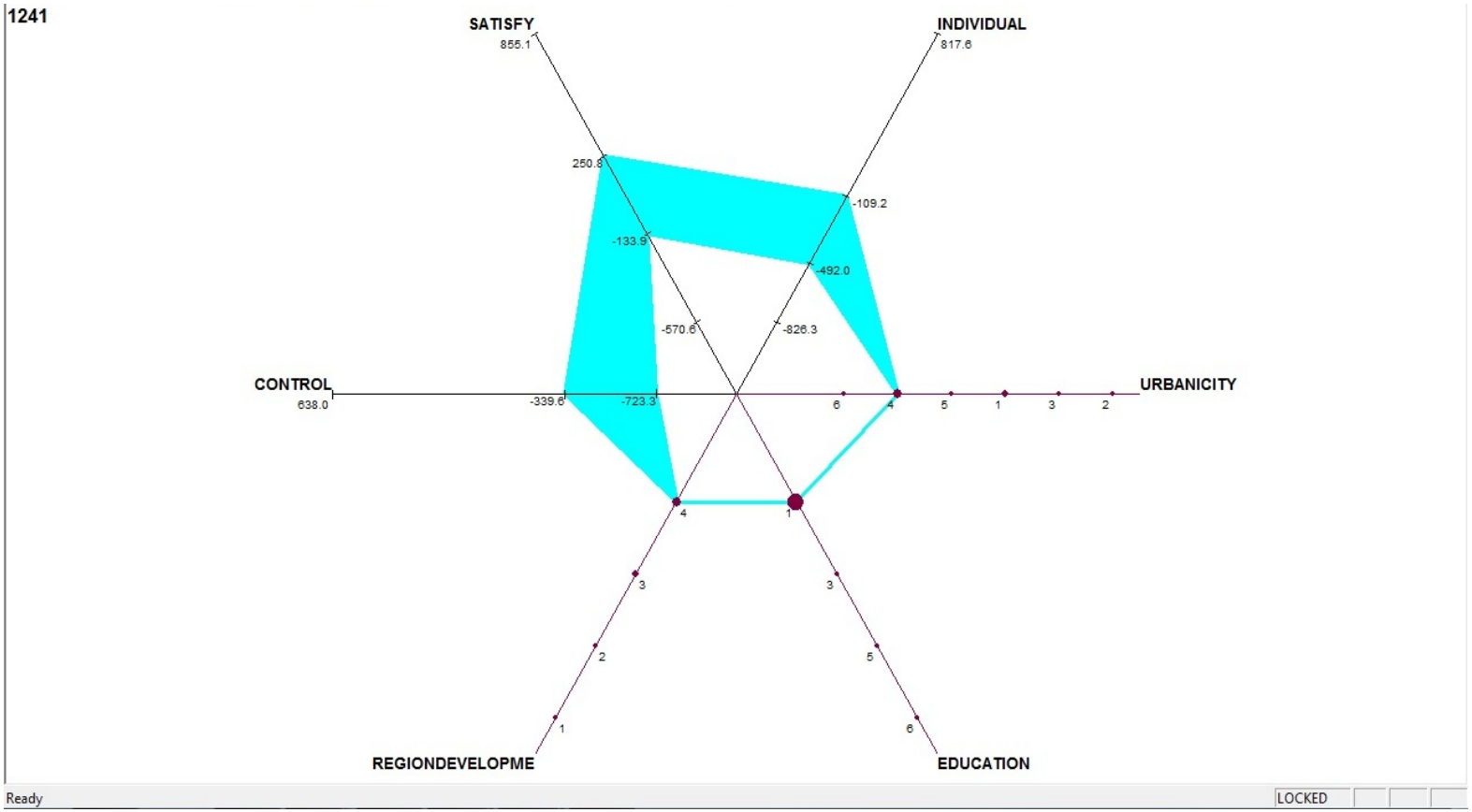
Set of 12 objects was built (2 combinations not observed).

Coding: 1241 $\rightarrow$ Gender $= 1$, Age ≤ 24 and Urban $= 1$

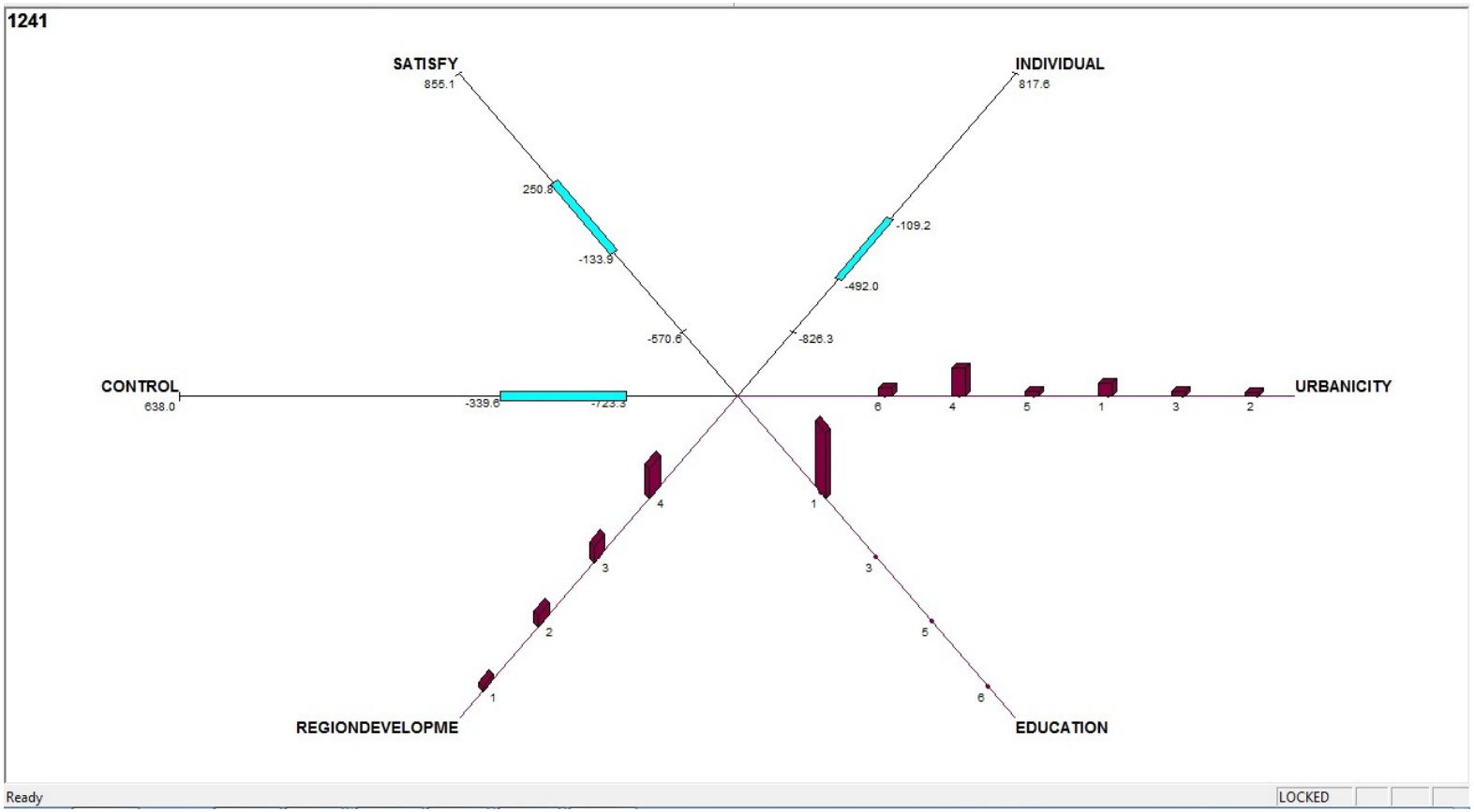
SODAS: Visualization

	URBANICITY	INCOMELEVEL	EDUCATION	REGIONDEVELOPME	CONTROL	SATISFY	INDIVIDUAL	WELFARE
1241	6 (0.14), 4 (0.43), 5 (0.08), 1 (0.21), 3 (0.09), 2 (0.06)	25 (0.91), 75 (0.02), 50 (0.07)	1 (1.00)	4 (0.47), 3 (0.25), 2 (0.19), 1 (0.09)	[-723.25 : -339.65]	[-133.91 : 250.83]	[-492.03 : -109.17]	[-229.95 : 184.4
1242	6 (0.08), 4 (0.49), 5 (0.08), 1 (0.15), 3 (0.12), 2 (0.07)	25 (0.56), 75 (0.15), 50 (0.29)	3 (0.98), 5 (0.02)	4 (0.49), 3 (0.32), 2 (0.08), 1 (0.10)	[-243.87 : 203.99]	[-134.15 : 314.80]	[-658.77 : -237.90]	[86.57 : 515.3
1441	6 (0.13), 4 (0.41), 5 (0.09), 1 (0.24), 3 (0.09), 2 (0.04)	25 (0.16), 75 (0.32), 50 (0.18), 90 (0.22), 100 (0.13)	1 (0.27), 3 (0.73)	4 (0.56), 3 (0.22), 2 (0.18), 1 (0.04)	[-195.45 : 90.10]	[-75.00 : 191.17]	[-592.28 : -299.04]	[-138.88 : 138.2
1442	6 (0.18), 4 (0.50), 5 (0.11), 1 (0.15), 3 (0.03), 2 (0.03)	25 (0.12), 75 (0.18), 50 (0.07), 90 (0.28), 100 (0.35)	5 (0.55), 6 (0.45)	4 (0.64), 3 (0.21), 2 (0.14), 1 (0.01)	[-279.12 : 120.61]	[342.58 : 705.03]	[-826.25 : -484.48]	[-103.32 : 306.7
1641	6 (0.12), 4 (0.38), 5 (0.08), 1 (0.30), 3 (0.10), 2 (0.02)	25 (0.11), 75 (0.27), 50 (0.27), 90 (0.24), 100 (0.13)	1 (0.50), 3 (0.50)	4 (0.46), 3 (0.24), 2 (0.19), 1 (0.11)	[-220.14 : 54.11]	[57.13 : 333.19]	[-88.31 : 161.99]	[-429.87 : -126.
1642	6 (0.10), 4 (0.57), 5 (0.10), 1 (0.10), 3 (0.09), 2 (0.04)	25 (0.04), 75 (0.13), 50 (0.06), 90 (0.22), 100 (0.55)	5 (0.43), 6 (0.57)	4 (0.57), 3 (0.24), 2 (0.17), 1 (0.01)	[-182.92 : 241.02]	[269.64 : 685.19]	[-514.62 : -24.76]	[-185.49 : 238.6
1741	6 (0.10), 4 (0.39), 5 (0.10), 1 (0.31), 3 (0.06), 2 (0.04)	25 (0.14), 75 (0.20), 50 (0.61), 90 (0.04), 100 (0.01)	1 (0.83), 3 (0.17)	4 (0.45), 3 (0.24), 2 (0.15), 1 (0.15)	[-158.77 : 320.11]	[52.89 : 504.42]	[192.71 : 584.90]	[-1117.93 : -607.
2241	6 (0.19), 4 (0.47), 5 (0.09), 1 (0.14), 3 (0.08), 2 (0.03)	25 (0.99), 75 (0.01)	1 (1.00)	4 (0.50), 3 (0.25), 2 (0.15), 1 (0.10)	[-431.93 : -43.55]	[-570.57 : -259.45]	[-240.45 : 136.85]	[62.28 : 429.7
2242	6 (0.22), 4 (0.42), 5 (0.13), 1 (0.15), 3 (0.03), 2 (0.04)	25 (0.73), 75 (0.04), 50 (0.21), 90 (0.01)	3 (0.93), 5 (0.07)	4 (0.61), 3 (0.22), 2 (0.12), 1 (0.04)	[-330.75 : 81.51]	[-522.62 : -62.18]	[-163.81 : 237.74]	[205.26 : 576.2
2441	6 (0.18), 4 (0.45), 5 (0.11), 1 (0.17), 3 (0.04), 2 (0.05)	25 (0.22), 75 (0.41), 50 (0.30), 90 (0.06), 100 (0.01)	1 (0.19), 3 (0.81)	4 (0.54), 3 (0.25), 2 (0.13), 1 (0.08)	[-86.03 : 172.79]	[-364.68 : -100.65]	[190.05 : 446.54]	[210.84 : 448.0
2442	6 (0.18), 4 (0.43), 5 (0.09), 1 (0.17), 3 (0.06), 2 (0.06)	25 (0.13), 75 (0.30), 50 (0.28), 90 (0.22), 100 (0.07)	5 (0.62), 6 (0.38)	4 (0.63), 3 (0.16), 2 (0.16), 1 (0.06)	[55.83 : 370.22]	[-450.05 : -138.20]	[-160.68 : 128.60]	[204.52 : 461.2
2641	6 (0.11), 4 (0.44), 5 (0.10), 1 (0.23), 3 (0.07), 2 (0.06)	25 (0.11), 75 (0.39), 50 (0.37), 90 (0.11), 100 (0.02)	1 (0.52), 3 (0.48)	4 (0.52), 3 (0.25), 2 (0.13), 1 (0.10)	[-9.01 : 269.21]	[-326.11 : -54.13]	[296.88 : 547.72]	[-172.77 : 65.3
2642	6 (0.23), 4 (0.41), 5 (0.13), 1 (0.11), 3 (0.10), 2 (0.02)	25 (0.05), 75 (0.30), 50 (0.10), 90 (0.35), 100 (0.20)	5 (0.55), 6 (0.45)	4 (0.72), 3 (0.16), 2 (0.06), 1 (0.05)	[133.40 : 478.04]	[-219.20 : 188.25]	[-275.80 : 70.06]	[-84.06 : 173.0
2741	6 (0.14), 4 (0.41), 5 (0.08), 1 (0.26), 3 (0.07), 2 (0.03)	25 (0.21), 75 (0.07), 50 (0.69), 90 (0.02)	1 (0.81), 3 (0.19)	4 (0.39), 3 (0.25), 2 (0.29), 1 (0.06)	[-9.00 : 389.63]	[-286.07 : 109.07]	[484.35 : 817.59]	[-623.18 : -179.

SODAS: Zoom Star 2D



SODAS: Zoom Star 3D



Outline

SODAS: Importing "Native" data

Symbolic "native data : Data are originally symbolic.
SODAS allows for the importation of :

- Quantitative single variables;
- Categorical single variables;
- Categorical multi-valued symbolic variables;
- Interval symbolic variables;
- Modal symbolic variables.

Importation to SODAS :

Record the data in a classical file: ASCII format.

- Interval data: one colunm for the minimum, one for the maximum
- Distributional data: one column per category

SODAS: Aggregating data with DB2SO

It is assumed that a set of individuals is stored in a database and distributed into some groups.
DB2SO builds one symbolic description for each group of individuals.

- + connection to a database
- + retrieving individuals distributed into groups by a SQL query
- optionally defining dependencies between variables
- optionally adding single-valued variables, e.g., **class variable**
- optionally adding multi-valued variables
- optionally adding taxonomies on variable domains
- optionally simplifying generated descriptions using the reduction facility
- specifying exportation and visualisation format of variables
- visualising all work that you have already done
- exporting generated descriptions to a SODAS file
- saving the current session to be able to restart it later

SODAS: Aggregating data with DB2SO

Query type I

- Record your microdata in a data base, e.g., ACCESS
- Build a query
- The query must contain at least three columns:
  - 1st column : individual identification
  - 2nd column : group identification
  - next columns : original variables

Other possibilities of aggregation

- Record data in a Excel or SPSS file
- Numerical variables:
  - Determine Minima and Maxima for interval-valued variables
  - Determine Quantiles for histogram-valued variables
- Categorical variables:
  - Determine lists for categorical multi-valued variables
  - Determine category frequencies for categorical modal variables

Outline

Analysis Issues

To represent data taking into account internal variability within each observation, variables have been allowed to assume new forms.

- Are we still in the same framework when we allow for the variables to take multiple values?
- Are the definitions of basic statistical notions still so straightforward?
- What properties remain valid?

Analysis Issues

Quantitative variables:

- Evaluation of dispersion $\rightarrow$ consequences in the design of multivariate methods
- Clustering $\rightarrow$ standardization: different standardization techniques for interval-valued variables proposed
- Many methodologies rely on linear combinations and on the properties of dispersion measures under linear transformations
- How to define a linear combination of symbolic variables ?
- Linear combinations of interval-valued variables: discussed in (Duarte Silva & Brito, Comp. Stat., 2006)

Analysis Issues

- Different approaches considered by various authors
- Most existing methods for the analysis of such data rely on a non-parametric descriptive approach
- Recent work with parametric models - Brito & Duarte Silva, 2012 ; Neto *et al*, 2011.