From Micro-Data to Macro-Data: Symbolic Data Analysis

Paula Brito

Fac. Economics & LIAAD-INESC TEC, University of Porto, PORTUGAL

European Conference on Quality in Official Statistics (Q2014)
Vienna, June $2^{nd} - 5^{th}$ 2014

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 Data Analysis (SDA) approach :
to take into account **variability** inherent to the data

Variability occurs when we have

- Data about students, but : analyse the schools - not the students
- Data about attendants, but : analyse the cultural events - not each individual attendant
- Data about purchases, but : analyse the clients (or classes of clients) - not the individual purchases
- Data about people, but : analyse the parishes, the cities, sociological groups - 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 cultural events
(e.g. museum exhibitions, theatre/cinema festival,...)

Event	Age category	Job group	Salary	Event evaluation
A	[15, 40]	{manager (0.20), clerk (0.30), scien-lib (0.1), student (0.40)}	$\{[0, 1.5[\,, 0.25; [1.5, 2.5[\,, 0.45; [2.5, 4[\,, 0.25; \geq 4, 0.05\}$	$\{1, 0.05; 2, 0.30; 3, 0.40; 4, 0.15; 5, 0.10\}$
B	[25, 55]	{manager (0.20), clerk (0.25), scien-lib (0.4), student (0.15)}	$\{[0, 1.5[\,, 0.15; [1.5, 2.5[\,, 0.35; [2.5, 4[\,, 0.30; \geq 4, 0.20\}$	$\{1, 0.05; 2, 0.25; 3, 0.30; 4, 0.25; 5, 0.15\}$
C	[12, 70]	{manager (0.20), clerk (0.35), scien-lib (0.20), student (0.25)}	$\{[0, 1.5[\,, 0.20; [1.5, 2.5[\,, 0.40; [2.5, 4[\,, 0.30; \geq 4, 0.10\}$	$\{1, 0.10; 2, 0.20; 3, 0.35; 4, 0.25; 5, 0.10\}$

Sources of symbolic data

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

Sources of symbolic data: Aggregation of micro-data

Name	Ammount	Event	Payement
A	5	cinema	Cash
A	25	concert	Visa
B	20	theatre	Electron
A	15	concert	Cash
C	40	theatre	Visa
B	8	cinema	Electron
A	10	museum	Cash
C	30	concert	Mastercard
...	...	...	...

Temporal aggregation $\downarrow$

Name	Ammount	Event	Payement
A	[5, 25]	{cinema(1/3), theatre(0), concert(1/3), museum(1/3)}	{Cash, Visa}
B	[8, 20]	{cinema(1/2), theatre(1/2), concert(0), museum(0)}	{Electron}
C	[30, 40]	{cinema(0), theatre(1/2), concert(1/2), museum(0)}	{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 $\downarrow$

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]

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

Interval data

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

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	$[10, 20]$

Age of attendants in cultural events :

	Age Attendants
Event A	$[15, 40]$
Event B	$[25, 55]$
Event C	$[12, 70]$

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

$\downarrow$

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	$[17, 100]$	$[0.8, 31.7]$	$[0.05, 6.57]$	$[0.09, 7.65]$

Distribution-Valued Data

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

Event	Job category	Event evaluation
A	{manager (0.20), clerk (0.30), scien-lib (0.1), student (0.40)}	$\{1, 0.05; 2, 0.30; 3, 0.40; 4, 0.15; 5, 0.10\}$
B	{manager (0.20), clerk (0.25), scien-lib (0.4), student (0.15)}	$\{1, 0.05; 2, 0.25; 3, 0.30; 4, 0.25; 5, 0.15\}$
C	{manager (0.20), clerk (0.35), scien-lib (0.20), student (0.25)}	$\{1, 0.10; 2, 0.20; 3, 0.35; 4, 0.25; 5, 0.10\}$

Histogram-valued variables: Example

Studying the performance of some administrative offices - time people have to wait before being taken care of:

Office	Waiting Times (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

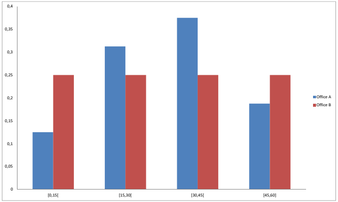
Average waiting time : 29.0 minutes for both offices

Description in terms of histograms :

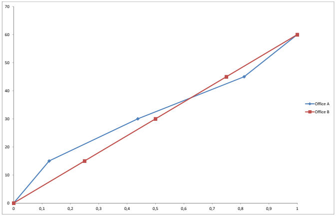
Office	Waiting Times (minutes)
A	$\{[0, 15[, 0.125; [15, 30[, 0.3125; [30, 45[, 0.375; [45, 60], 0.1875\}$
B	$\{[0, 15[, 0.25; [15, 30[, 0.25; [30, 45[, 0.25; [45, 60], 0.25\}$

Histogram-valued variables: Example

Histograms :



Quantile functions :



Histogram-valued variables: Example

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

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
- ...

Methods for multivariate data analysis

- Interval-valued variables: a special case of histogram-valued variables
- Methods first developed for interval-valued variables:
- Greater effort in addressing and designing methods for interval data

A large number of methods have to this day been developed for multivariate analysis, including :

- Clustering - Partitioning (crisp, fuzy), Hierarchical, SOM,...
- Classification - LDA, Decision tres, Neural networks,...
- Factorial analysis - PCA, Generalized canonical analysis
- Multiple Regression
- Time series analysis

Outline

Analyzing Unemployment Data: Methodology

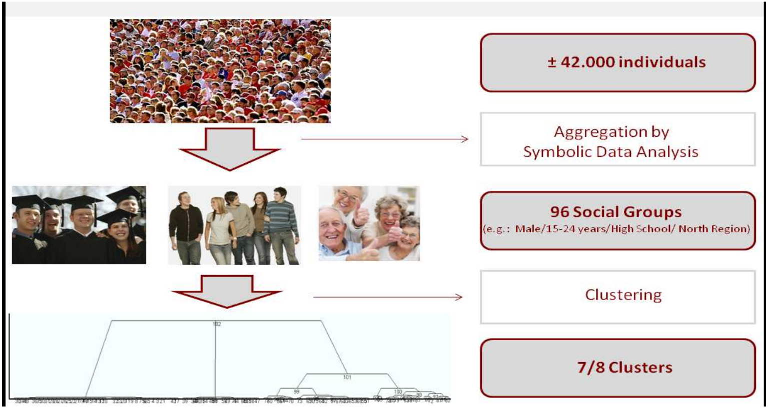
Using a symbolic data analysis approach

- we aggregated the micro-data from the Employment Survey
- in social groups based on age, gender and education
- obtaining 96 social groups

Cluster analysis

- Tranversally
- Through time

Methodology



Variables

The descriptive variables are :

- Unemployment time : Numerical $\longrightarrow$ Interval-valued
- Satisfaction : Categorical $\longrightarrow$ Modal categorical
- Sector : Categorical $\longrightarrow$ Modal categorical
- Job situation : Categorical $\longrightarrow$ Modal categorical
- Employment situation : Categorical $\longrightarrow$ Modal categorical
- Second activity : Categorical $\longrightarrow$ Modal categorical
- Revenue source : Categorical $\longrightarrow$ Modal categorical
- Reason for abandon : Categorical $\longrightarrow$ Modal categorical

The data

VSTAR - IE2008_1T.sds

	tempo_actividad	Tempo_desempreg	Fonte_de_rendim	
Masculino/>=65/Básic	[30.00 : 86.00]	[24.00 : 24.00]	(0.95), Depen (0.00), Salár (0.01), Lucro (0.03), Espec (0.00), Subsi (0.00), Rendi	Refor (0.46), Outra (0.01), Doenç (0.28), Refor (0.15), Outra
Feminino/45-64/Básic	[0.00 : 57.00]	[1.00 : 372.00]	(0.31), Lucro (0.07), Espec (0.05), Subsi (0.05), Rendi (0.02), Subsi (0.01), Tr Pa	Refor (0.02), Outra (0.11), Doenç (0.32), Refor (0.08),
Feminino/>=65/Básico	[8.00 : 88.00]	Missing Value	Refor (0.91), Depen (0.06), Salár (0.00), Lucro (0.02), Espec (0.01), Tr Pa (0.00)	Refor (0.31), Outra (0.07), Doenç (0.40), Refor (0.10),
Feminino/45-64/Super	[8.00 : 53.00]	Missing Value	Refor (0.23), Depen (0.01), Salár (0.72), Lucro (0.04)	Refor (0.65), Outra (
Feminino/>=65/Superi	[35.00 : 64.00]	Missing Value	Refor (0.88), Salár (0.04), Lucro (0.08)	Refor (0.70),
Masculino/45-64/Bási	[10.00 : 58.00]	[0.00 : 159.00]	(0.52), Lucro (0.12), Espec (0.02), Subsi (0.05), Rendi (0.01), Subsi (0.01), Tr Pa	Refor (0.07), Outra (0.01), Doenç (0.33), Refor
Feminino/25-44/Básic	[0.00 : 35.00]	[1.00 : 325.00]	, Salár (0.61), Lucro (0.06), Espec (0.03), Subsi (0.03), Rendi (0.03), Subsi (0.01),	Outra (0.11), Doenç (0.15), Refor (0.00), Outra (0.02),
Masculino/15-24/Secu	[0.00 : 11.00]	[1.00 : 15.00]	Refor (0.00), Depen (0.62), Salár (0.35), Lucro (0.01), Subsi (0.01)	
Masculino/15-24/Supe	[0.00 : 6.00]	Missing Value	Depen (0.40), Salár (0.55), Lucro (0.05)	
Masculino/45-64/Secu	[15.00 : 53.00]	[9.00 : 35.00]	Refor (0.15), Depen (0.03), Salár (0.65), Lucro (0.11), Subsi (0.06)	Refor (0.08), Doenç (0.16), Refor
Feminino/45-64/Secun	[7.00 : 46.00]	[0.00 : 123.00]	Refor (0.19), Depen (0.13), Salár (0.58), Lucro (0.05), Subsi (0.04)	Refor (0.05), Outra (0.13), Doenç (0.15), Refor (0.30),
Masculino/25-44/Bási	[1.00 : 37.00]	[1.00 : 97.00]	(0.75), Lucro (0.09), Espec (0.01), Subsi (0.02), Rendi (0.00), Subsi (0.00), Tr Pa	Outra (0.01), Doenç (0.26), Refor (0.02), Outra (0.04)
Feminino/<15/Básico	Missing Value	Missing Value	Missing Value	
Feminino/<15//Norte	Missing Value	Missing Value	Missing Value	
Feminino/15-24/Básic	[0.00 : 12.00]	[1.00 : 61.00]	(0.01), Depen (0.69), Salár (0.28), Subsi (0.00), Rendi (0.01), Subsi (0.00), Tr Pa	Outra (0.15), Outra (0.05),
Feminino/15-24/Secun	[0.00 : 8.00]	[0.00 : 10.00]	Refor (0.01), Depen (0.72), Salár (0.26), Subsi (0.00)	Estud
Feminino/25-44/Secun	[0.00 : 28.00]	[2.00 : 88.00]	(0.01), Depen (0.18), Salár (0.73), Lucro (0.04), Subsi (0.03), Rendi (0.01), Subsi	Outra (0.05), Doenç (0.08), Outra (0.11),
Feminino/15-24/Super	[0.00 : 10.00]	[46.00 : 46.00]	Depen (0.36), Salár (0.64)	
Masculino/15-24/Bási	[0.00 : 18.00]	[1.00 : 33.00]	(0.01), Depen (0.57), Salár (0.41), Lucro (0.01), Subsi (0.00), Tr Pa (0.00), Ajuda	Doenç (0.10), Outra (0.15),
Feminino/25-44/Super	[0.00 : 26.00]	[2.00 : 80.00]	Depen (0.13), Salár (0.81), Lucro (0.03), Subsi (0.01), Rendi (0.00), Subsi (0.00),	Outra (0.09), Doenç (0.09),
Masculino/<15/Básico	Missing Value	Missing Value	Missing Value	
Masculino/25-44/Supe	[0.00 : 27.00]	[2.00 : 100.00]	Depen (0.09), Salár (0.79), Lucro (0.09), Subsi (0.01), Subsi (0.01)	Estud
Masculino/<15//Norte	Missing Value	Missing Value	Missing Value	
Masculino/45-64/Supe	[12.00 : 48.00]	[9.00 : 19.00]	Refor (0.12), Depen (0.01), Salár (0.76), Lucro (0.09), Subsi (0.02)	Refor (0.33),
Masculino/25-44/Secu	[0.00 : 32.00]	[1.00 : 46.00]	(0.16), Salár (0.74), Lucro (0.06), Espec (0.01), Subsi (0.02), Subsi (0.01), Tr Pa	Outra (0.06), Doenç (0.11),

Ready

Outline

Aggregating Surveys

Symbolic Data Analysis allows for the merging of different surveys in the same population:

- Aggregate the 1st survey on given variables, e.g. age, X gender
- Aggregate the 2nd survey on the **same** variables
- Therefore, the groups formed are the same
- Merge the two surveys : for each group formed, add the variables of one survey to the variables of the other survey

Aggregating Surveys

Symbolic Data Analysis allows for the merging of similar surveys from different populations - e.g., different countries:

- Aggregate the 1st survey on the population (e.g., country) and other given variables, e.g. age, X gender
- Aggregate the 2nd survey on the **same** variables
- Therefore, the groups formed in each population are the same
- Merge the two surveys : merge the groups of one survey with those of the other survey: add rows
- The variables are the same in both surveys, so nothing to do with the columns

Outline

1. Symbolic data
2. Analyzing Unemployment Data
3. Potential New Applications in Official Statistics
4. Conclusion
5. References

Concluding Remarks

Symbolic Data Analysis allows to

- Consider big data-sets
- Analyse data at the required level, keeping intrinsic variability information
- Use surveys from different years / countries
(not necessarily the same people !)

Outline

Books and Main Papers

- Bock, H.-H.; Diday, E. (2000): *Analysis of Symbolic Data: Exploratory methods for extracting statistical information from complex data.* Berlin-Heidelberg: Springer-Verlag.
- Billard, L., Diday, E. (2007): *Symbolic Data Analysis: Conceptual Statistics and Data Mining.* Wiley.
- Diday, E., Noirhomme-Fraiture, M. (2008): *Symbolic Data Analysis and the SODAS Software.* Wiley.
- Special Issue of *Statistical Analysis and Data Mining*. Vol. 4, Issue 2, April 2011. Wiley, on behalf of the American Statistical Association.
- Billard, L. and Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic Data Analysis. *Journal of the American Statistical Association*, 98 (462), pp. 470487.
- Noirhomme-Fraiture, M. and Brito, P. (2011). Far beyond the classical data models: Symbolic data analysis. *Statistical Analysis and Data Mining*, 4(2), 157–170.