Symbolic Data Analysis Of Complex Data: several directions of research

Edwin Diday
CEREMADE Paris Dauphine University

OUTLINE

- What are Complex data?
- What are “symbolic data”?
- How “Symbolic Data” are build?
- Symbolic Data are Complex data?
- From Complex Data to Symbolic Data
- What is “Symbolic Data Analysis” (SDA)?
- Open directions of research
- Conclusion: SDA gives a framework for Complex Data Analysis (CDA)

OUTLINE

- **What are Complex data?**
- What are “symbolic data”?
- How “Symbolic Data” are build?
- Symbolic Data are Complex data?
- From Complex Data to Symbolic Data
- What is “Symbolic Data Analysis” (SDA)?
- Open directions of research
- SDA gives a framework for Complex Data Analysis (CDA)

What are Complex data?

Any data which cannot be considered as a standard “observations x variables” data table.

Examples

- several data tables describing different kind of observations.
- Hierarchical Data
- Textual Data in each cell of the data table
- Time series Data in each cell .

OUTLINE

- What are Complex data?
- **What are “symbolic data”?**
- How “Symbolic Data” are build?
- Symbolic Data are Complex data?
- From Complex Data to Symbolic Data
- What is “Symbolic Data Analysis” (SDA)?
- Open directions of research
- SDA gives a framework for Complex Data Analysis (CDA)

What are “symbolic data”?

Any data taking care on the variation inside classes of standard observation.

- **each cell of the data table can contain:**
- **A number, a category, an interval, a sequence of categorical values, a sequence of weighted values , a Bar Chart, a histogram, a distribution, …**

Example of SYMBOLIC DATA

TEAM OF THE FRENCH CUP	WEIGHT	NATIONALITY	NB OF GOALS
MARSEILLES	[75 , 89]	{French}	{0.8 (0), 0.2 (1)}
LYON	[80, 95]	{Fr, Alg, Arg }	{0.1 (0), 0.3 (1), …}
PARIS-ST G.	[76, 95]	{Fr, Tun }	{0.4 (0), 0.2 (1), …}
NANTES	[70, 85]	{Fr, Engl, Arg }	{0.2 (0), 0.5 (1), …}

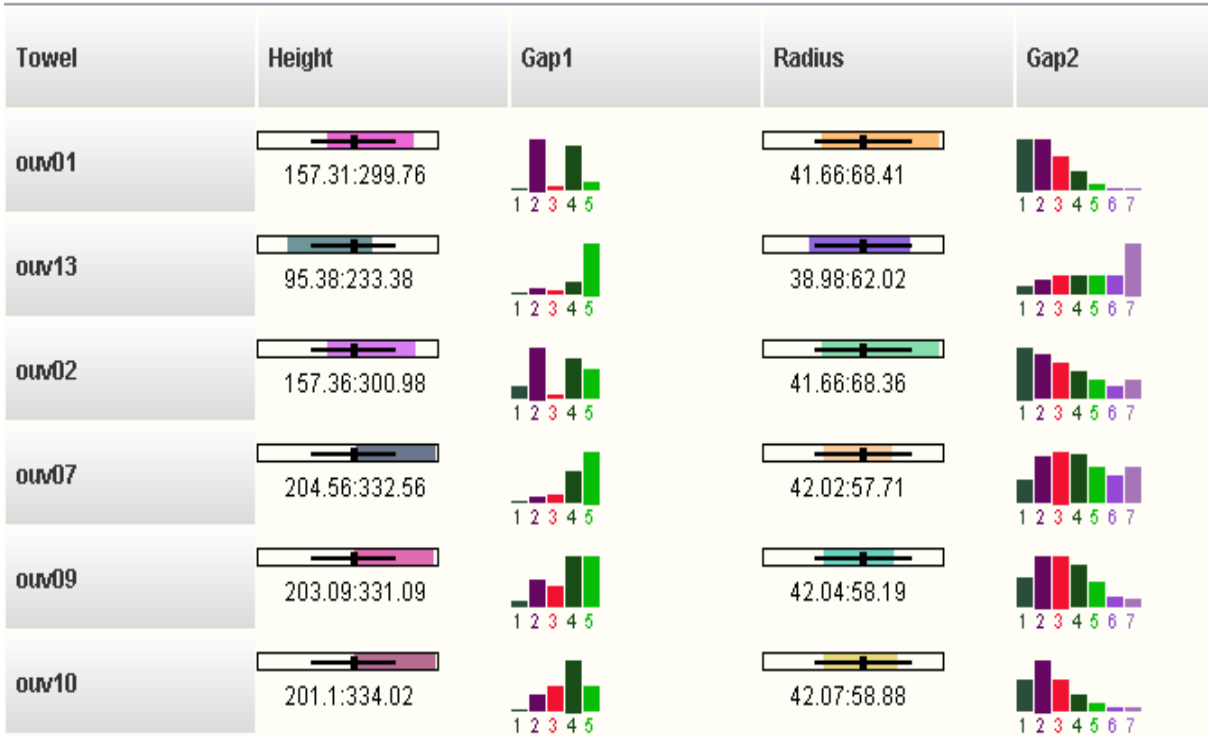
Here the variation (of weight, nationality, …) concerns the players of each team.

Therefore each cell can contain:

A number, an interval, a sequence of categorical values, a sequence of weighted values as a histogram, a distribution, …

THIS NEW KIND OF VARIABLES ARE CALLED « SYMBOLIC » BECAUSE THEY ARE NOT PURELY NUMERICAL IN ORDER TO EXPRESS THE INTERNAL VARIATION INSIDE EACH CONCEPT.

SYMBOLIC DATA TABLE SOFTWARE*

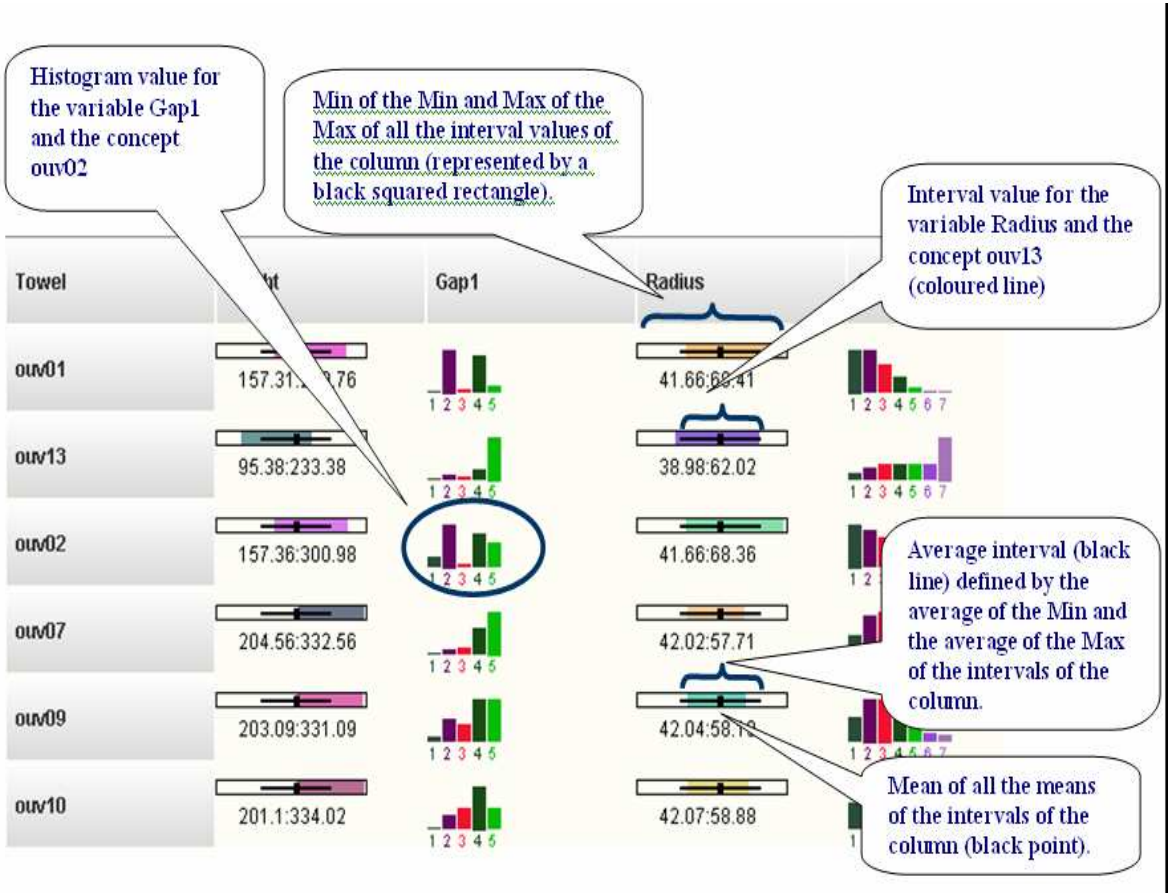


Towel	Height	Gap1	Radius	Gap2
ouv01	157.31:299.76	1 2 3 4 5	41.66:68.41	1 2 3 4 5 6 7
ouv13	95.38:233.38	1 2 3 4 5	38.98:62.02	1 2 3 4 5 6 7
ouv02	157.36:300.98	1 2 3 4 5	41.66:68.36	1 2 3 4 5 6 7
ouv07	204.56:332.56	1 2 3 4 5	42.02:57.71	1 2 3 4 5 6 7
ouv09	203.09:331.09	1 2 3 4 5	42.04:58.19	1 2 3 4 5 6 7
ouv10	201.1:334.02	1 2 3 4 5	42.07:58.88	1 2 3 4 5 6 7

Scoring rows by min, max of intervals or frequencies or barchart is possible.

* SYROKKO Company eliezer@syrokko.com

SYMBOLIC DATA TABLE SOFTWARE*



Scoring variables is also possible in order to select the most discriminate variables of the rows

* SYROKKO Company eliezer@syrokko.com

OUTLINE

- What are Complex data?
- What are “symbolic data”?
- **How symbolic data are build?**
- Symbolic Data are Complex data?
- Complex data are Symbolic Data after transformation ?
- What is “Symbolic Data Analysis” (SDA)?
- SDA gives a framework for Complex Data Analysis (CDA)?
- Open directions of research.

First step: From Standard Data TABLE 1 To random variables in each cell TABLE 2

Table 1

	Y_1		Y_j	
w_1				
w_i			x_{ij}	
w_n				

In each cell a number or a category

Table 2

	X_1		X_j	
C_1				
C_i			$\mathbf{X_{ij}}$	
C_k				

In each cell a Random Variable $C_i \longrightarrow D_j$

Standard data table: w_i x Y_j

(Observations) x (Random Variables)

$Y_j(w_i) = \mathbf{x}_{ij}$ = a number or a category among Dj the domain of Y_j

Random Variables data table: C_i x X_j

(Classes of Observations) x (Random variables of random variable values)

$X_j\ (C_i) = \mathbf{X_{ij}}$ **is a random variable:**

$\mathbf{X_{ij}}\ (w) = \mathbf{x}$ = a number or a category

RANDOM VARIABLES FROM STANDARD DATA

Table1 : Standard data table

PLAYERS OF THE FRENCH CUP	WEIGHT Y_1	NATIONALITY Y_2	NB OF GOALS Y_3
ZIDANE	80	FRENCH	12

Table 2 : Random Variable data table

TEAM	WEIGHT X_1	NATIONALITY X_2	NB GOALS X_3
MARSEILLES	X_{11}	X_{12}	X_{13}
LYON	X_{21}	X_{22}	X_{23}
PARIS-ST G.	X_{31}	X_{32}	X_{33}
NANTES	X_{34}	X_{42}	X_{43}

➢**Table 1**: Here the variables Y_j (weight, nationality, …) are random variables defined on the players :

$Y_j: \Omega \rightarrow D_j$: $Y_1(\text{Zidane}) = 80$ ➡ Weight(Zidane) = 80

➢**Table 2**: here the variables Xj are random variables defined on the teams with Random Variable Xij as value:

➢$X_j(C_i) = X_{ij}$ where $X_{ij}: C_i \rightarrow D_j$ is a RV: $X_{i1}(\text{Zidane}) = 80$
Weight(Zidane) = 80 if Zidan belongs to C_i

SECOND STEP: From TABLE 2 (Random variable in each cell) To TABLE 3 (Symbolic Data in each cell)

Table 2

	X_1		X_j	
C_1				
C_i			$\mathbf{X_{ij}}$	
C_n				

In each cell a Random Variable $\mathbf{X_{ij}}$: $C_i \longrightarrow D_j$

Table 3

	Y'_1		Y'_j	
C_1				
C_i				
C_k				

In each cell a symbolic data

Random Variables data table: C_i x X_j

(Classes of Observations) x (Random variables of random variable values)

X_j (Ci) = $\mathbf{X_{ij}}$ **is a random variable:**

$\mathbf{X_{ij}}$ (w) = **x** = a number or a category

Symbolic Variables : C_i x Y'_j

(Classes of Observations) x (Random variables of Symbolic values)

Y'_j (Ci) = $\mathbf{H_{ij}}$ **is a barchart value or an interval representing variation inside class C_i for the variable Y_j .**

FROM RANDOM VARIABLE DATA TABLE 2 TO SYMBOLIC DATA TABLE 3

Standard case Table 1: the variables Y_j are random variables with numerical or categorical values.

Symbolic case Table 2: the variables X_j are Random Variables with Random Variables X_{ij} as values.

Symbolic data Table 3: the random variables X_{ij} are represented by:

. Probability densities

. Distributions

. Bar charts

. Inter-quartile intervals

. Parameters (mean square, standard déviation, …)

Symbolic Representation in TABLE 3 of the Random Variables of TABLE 2

TEAM OF THE FRENCH CUP	WEIGHT X_1	NATIONALITY X_2	NB OF GOALS X_3
MARSEILLES	[75 , 89]	{French}	{0.8 (0), 0.2 (1)}
LYON	[80, 95]	{Fr, Alg, Arg }	{0.1 (0), 0.3 (1), …}
PARIS-ST G.	[76, 95]	{Fr, Tun }	{0.4 (0), 0.2 (1), …}
NANTES	[70, 85]	{Fr, Engl, Arg }	{0.2 (0), 0.5 (1), …}

- In Table 2: the random variable X_1 is associated to "Weight", its value for the team "Lyon" is the Random Variable X_{21} defined on the players of Lyon with value a weight inside the Domain D_1 of possible weights.
- The representation of the Random Variable X_{12} is the interquartile interval: [80, 95]
- The representation of the Random Variable X_{13} is the Bar Chart: [0.8 (0), 0.2 (1)]

OUTLINE

- What are Complex data?
- What are “symbolic data”?
- How “Symbolic Data” are build?
- Symbolic Data are Complex data?
- Complex data are Symbolic Data after transformation ?
- What is “Symbolic Data Analysis” (SDA)?
- SDA gives a framework for Complex Data Analysis (CDA)?
- Open directions of research.

WHY SYMBOLIC DATA CANNOT BE REDUCED TO A CLASSICAL DATA TABLE?

Symbolic Data Table

Players category	Weight	Size	Nationality
Very good	[80, 95]	[1.70, 1.95]	{0.7 Eur, 0.3 Afr}

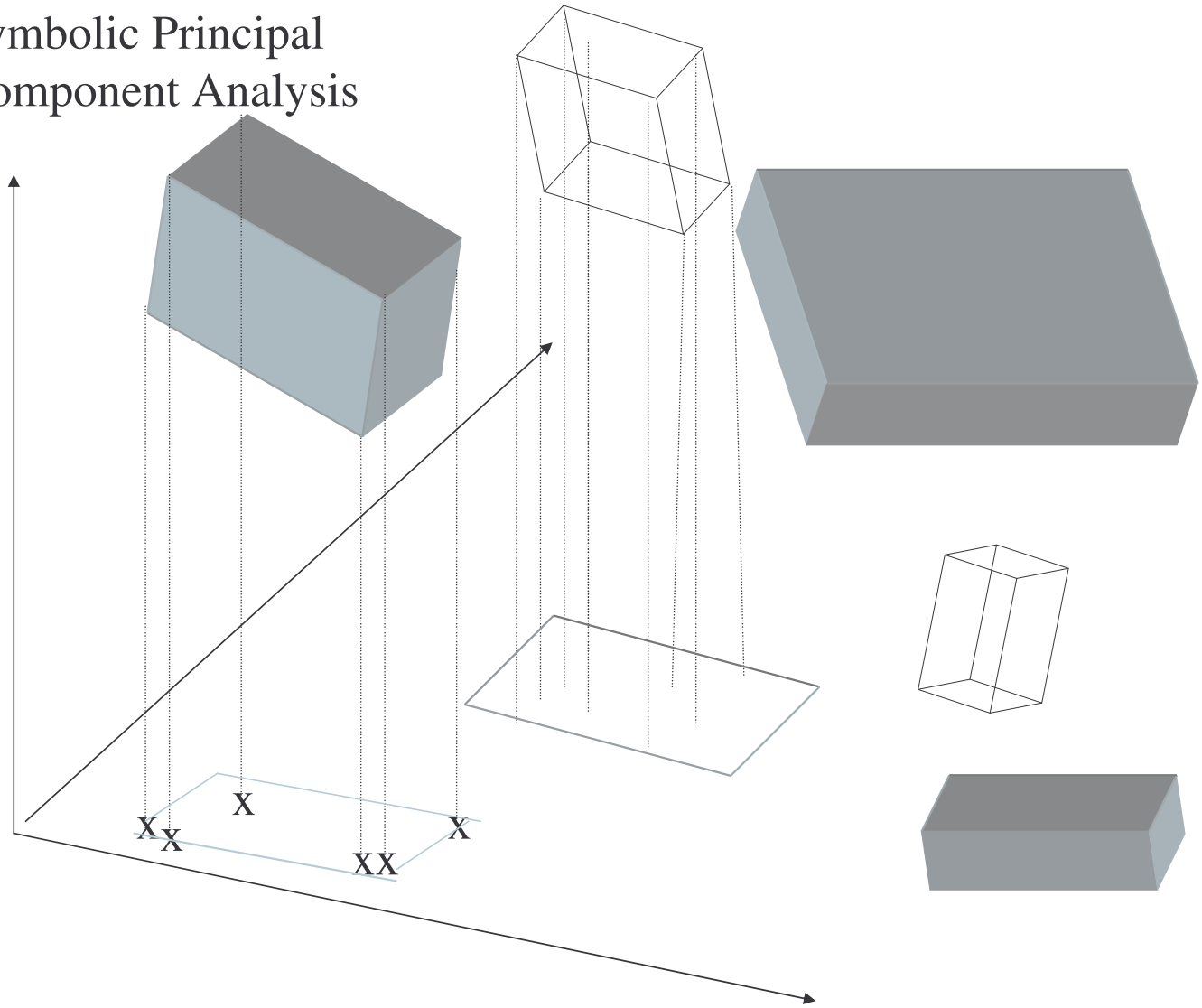
Transformation in classical data

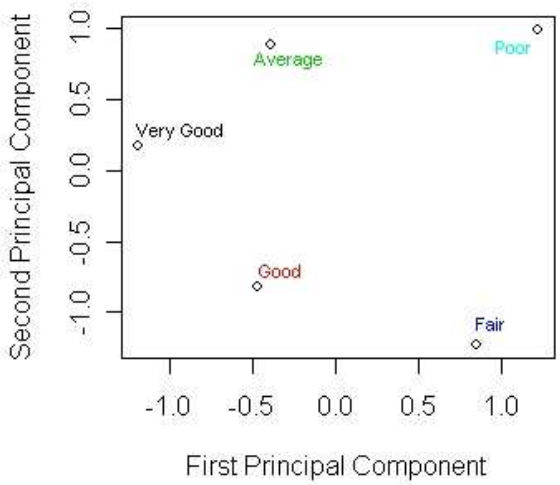
Players category	Weight Min	Weigh t Max	Size Min	Size Max	Eur	Afr
Very good	80	95	1.70	1.95	0. 7	0.3

Concern:

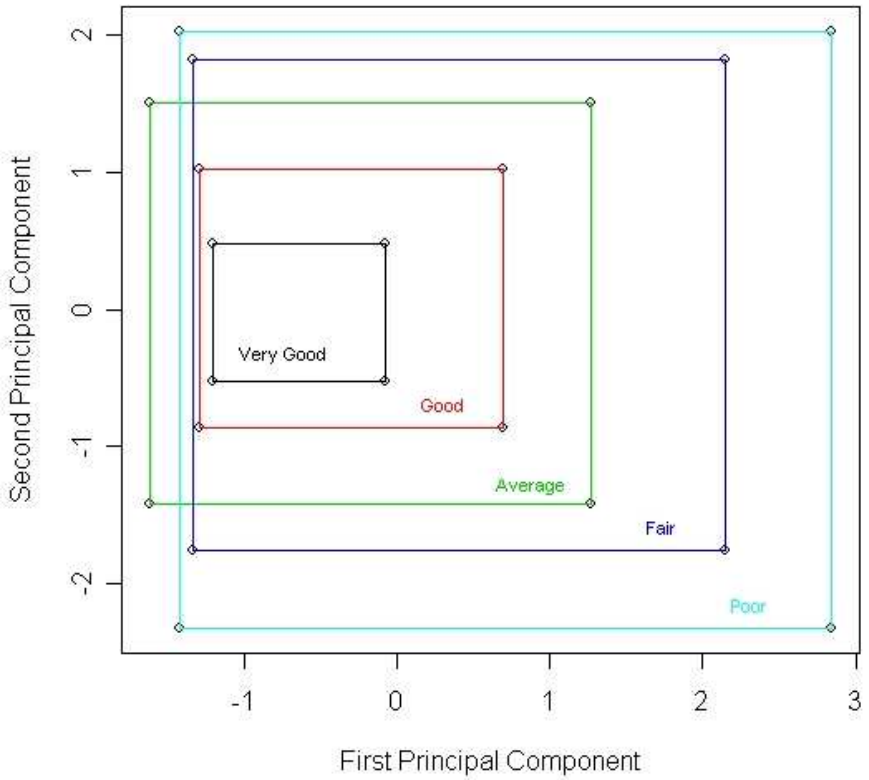
The initial variables are lost and the variation is lost!

Symbolic Principal Component Analysis



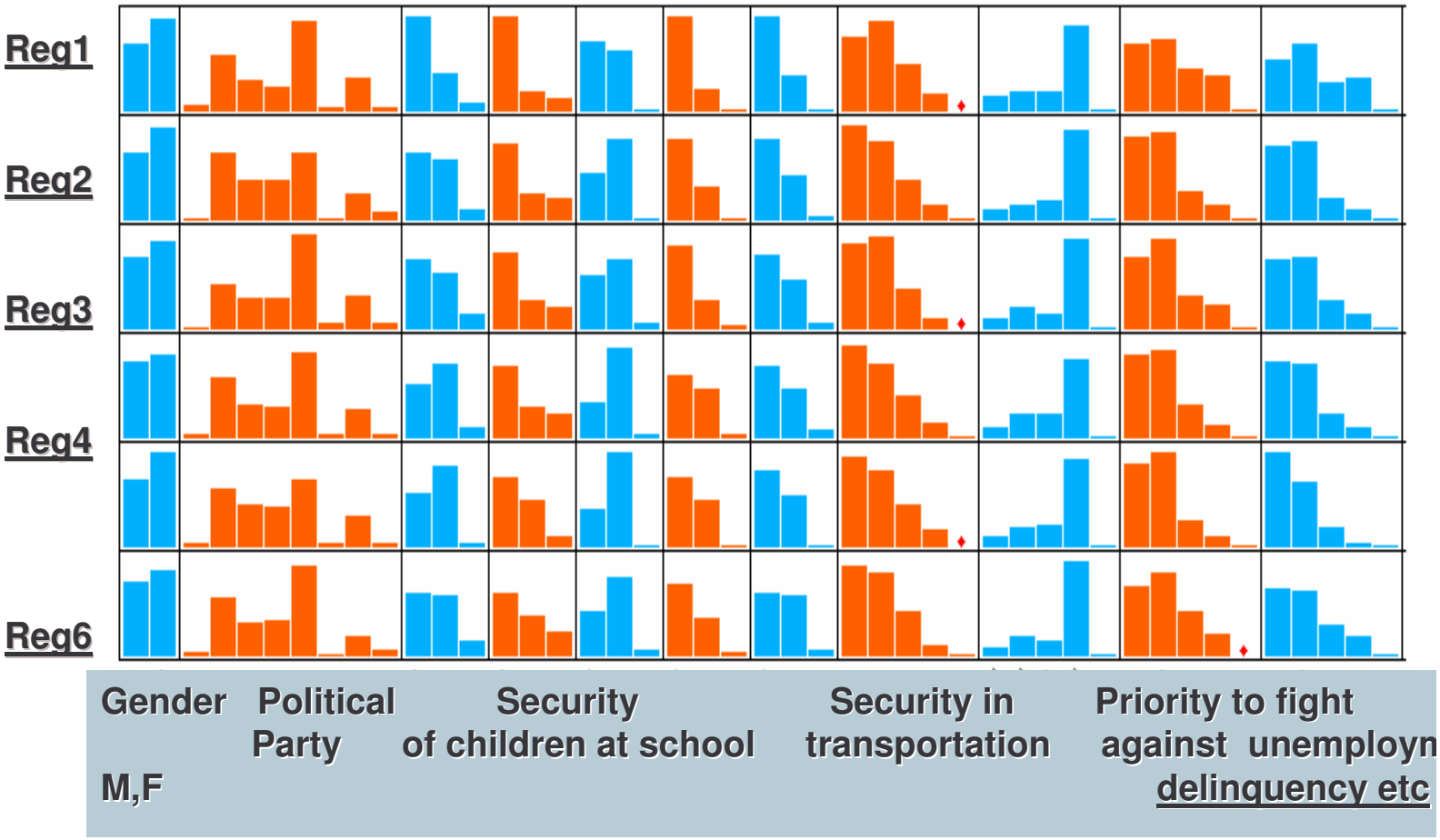


Classical Analysis
Loose variation

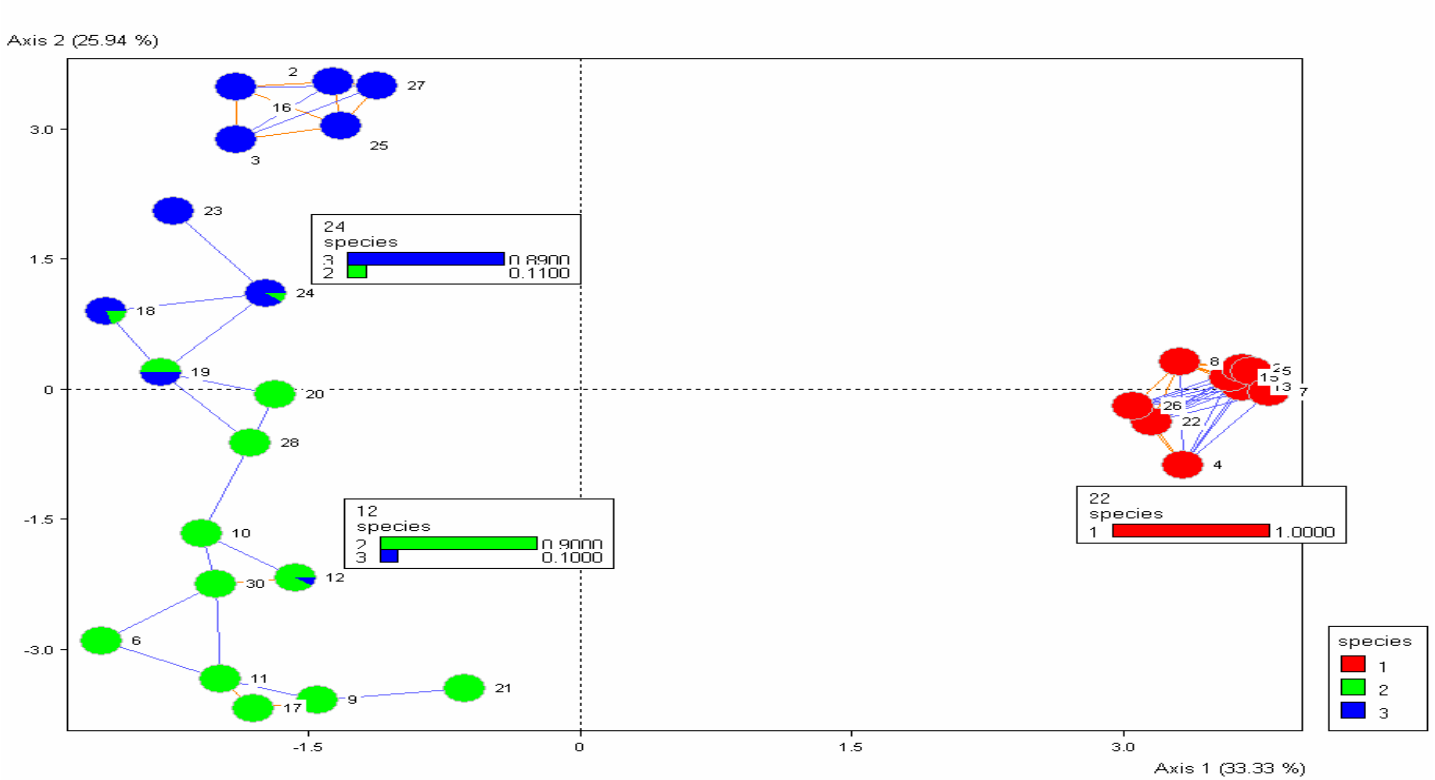


Symbolic Analysis
Take care of
variation

Tackle security problems in regions



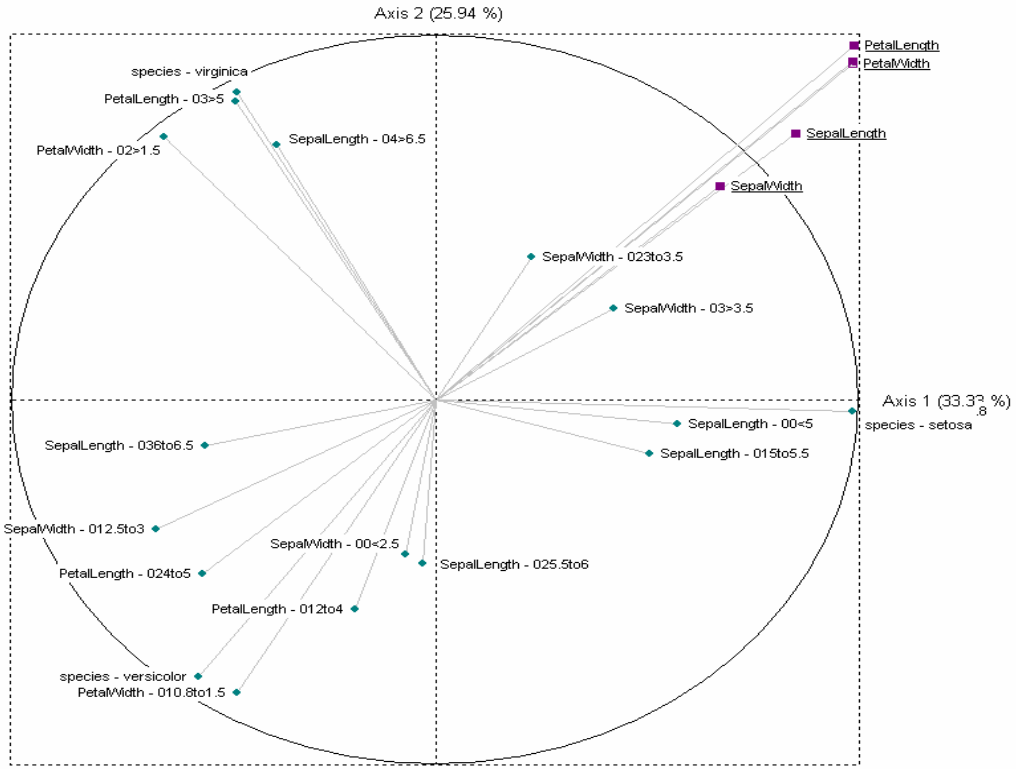
PCA and NETWORK OF BAR CHART DATA of 30 Iris Fisher Data Clusters*



Any symbolic variable can be projected. Here the species variable.

* SYROKKO Company eliezer@syrokko.com

The Symbolic Variables contributions are inside the smallest hyper cube containing the correlation circle of the bins



Conclusion:

Symbolic data are complex data as they cannot be reduced to standard data without loosing much information.

OUTLINE

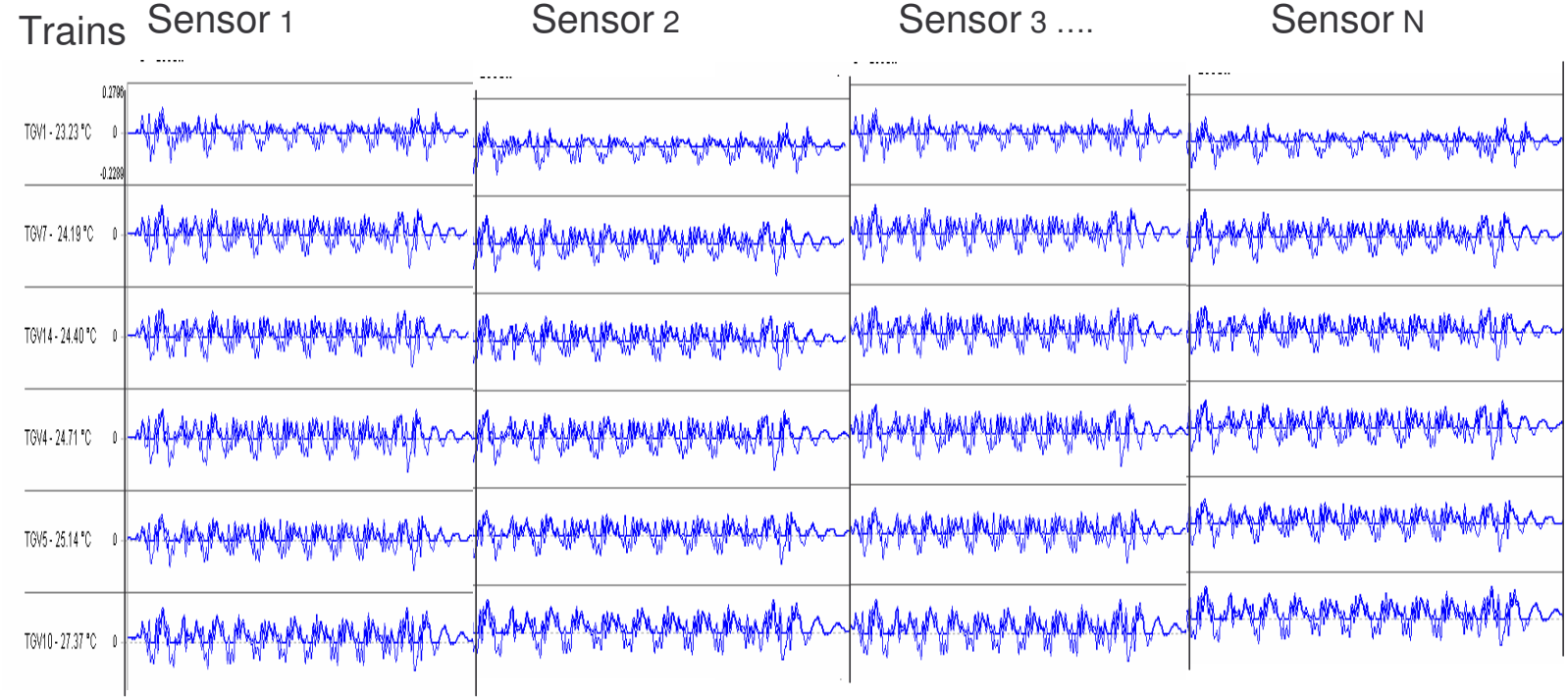
- What are Complex data?
- What are “symbolic data”?
- How “Symbolic Data” are build?
- Symbolic Data are Complex data?
- **From Complex data to Symbolic Data**
- What is “Symbolic Data Analysis” (SDA)?
- Open directions of research.
- SDA gives a framework for Complex Data Analysis (CDA)

Complex data are Symbolic Data ?

- Time series data table
- Multisource data tables
- Hierarchical data
- Textual Data
- Etc.

CAN BE TRANSFORMED IN SYMBOLIC DATA

Time series data table: Anomaly detection on a bridge
LCPC (Laboratoire Central Des Ponts et Chaussées and SNCF Data



Each row represents a train going on the bridge at a given temperature,

each cell contains until 800.000 values.

Each cell is transformed in HISTOGRAM from a PROJECTION or from WAVELETS

The symbolic aggregation approach

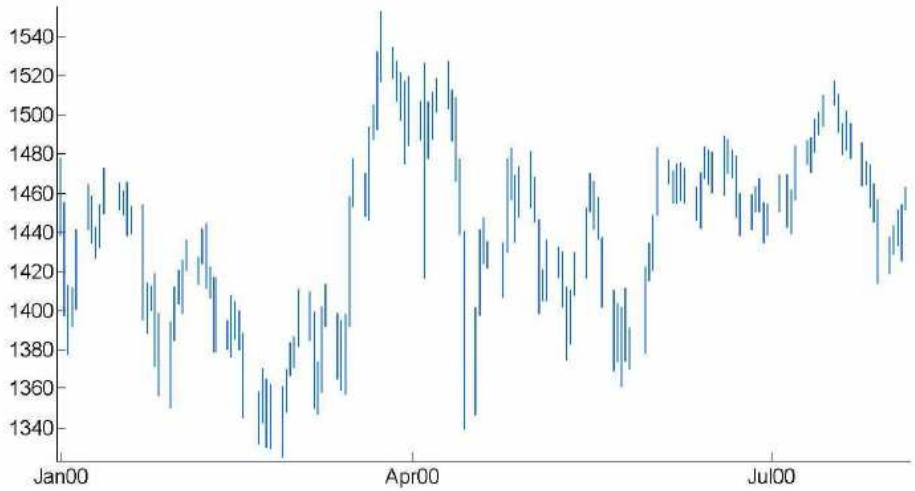


Figure 2: Interval time series of the high and low prices of the SP500 stock index.

Multisource data tables

FRANCE IS DIVIDED INTO 50 097 COUNTIES CALLED IRIS

IRIS are the level to study, initial data are confidential and multisource

Classical Data table

Household	IRIS	Size	Car Mark	SPC
Dupont	IRIS 55	2	Renault	3
Durand	IRIS 602	5	Renault	1
Boule	IRIS 498	3	Peugeot	2

Symbolic description of households in IRIS 1

IRIS	Size	Car Mark	SPC
IRIS 1	[0, 5]	Renault(43%), Citroën (21%)….	

Classical Data table

School	IRIS	TYPE
Condorcet	IRIS 605	Private
Laplace	IRIS 75	Public
Voltaire	IRIS 855	Public

Symbolic description of shools in IRIS 1

IRIS	TYPE	Spécialisation	
IRIS 1	{(private, 37%);(public, 63%)}	{(yes,17%); (no, 83%)}	

Concatenation

IRIS n = [Symb. Description of households] $\wedge$ [Symb. Description of School

NEW DATA: in one SYMBOLIC DATA TABLE describing each IRIS.

Multisource data tables

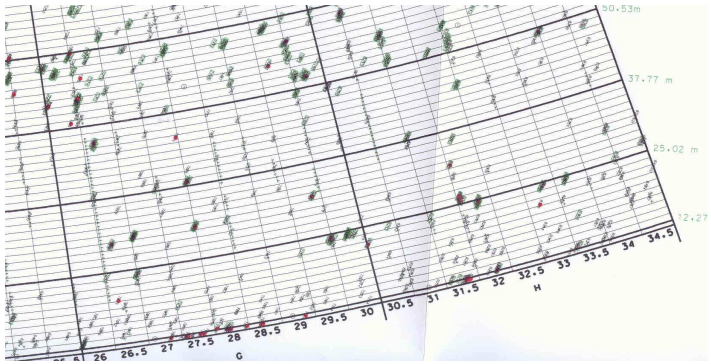
NUCLEAR POWER PLANT Nuclear thermal power station

Inspection :

Inspection machine

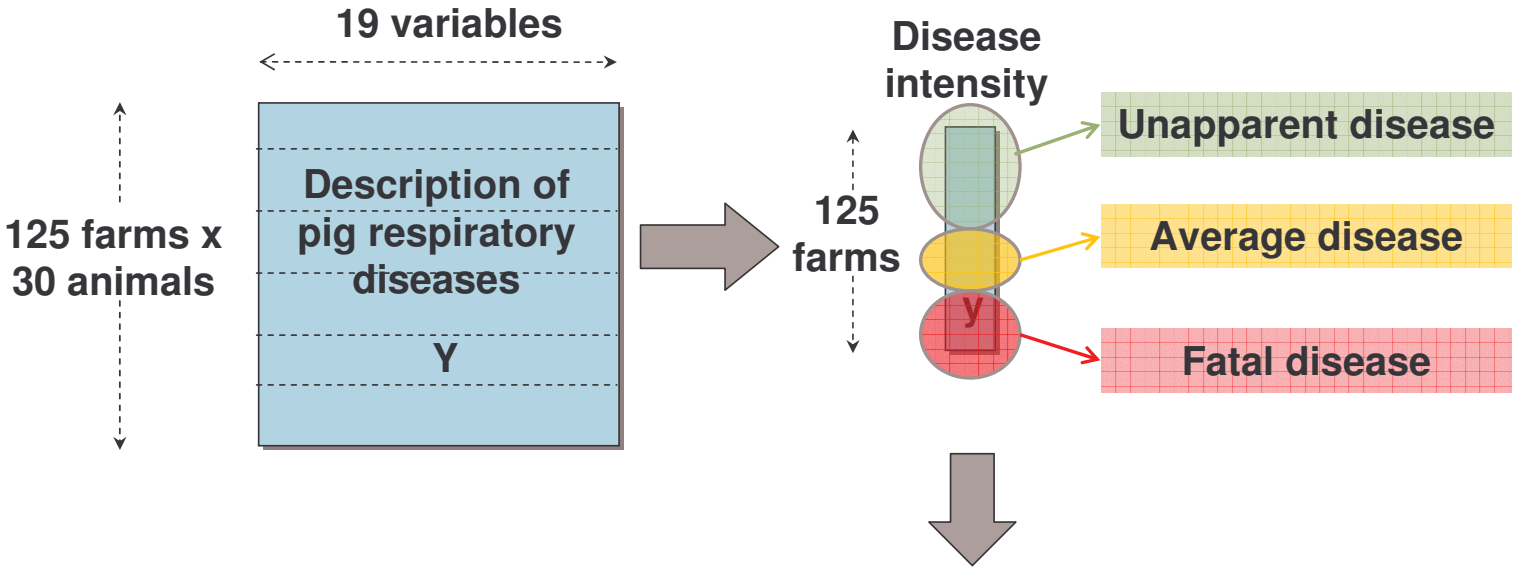
Craks

Cartography of the towel by a grid



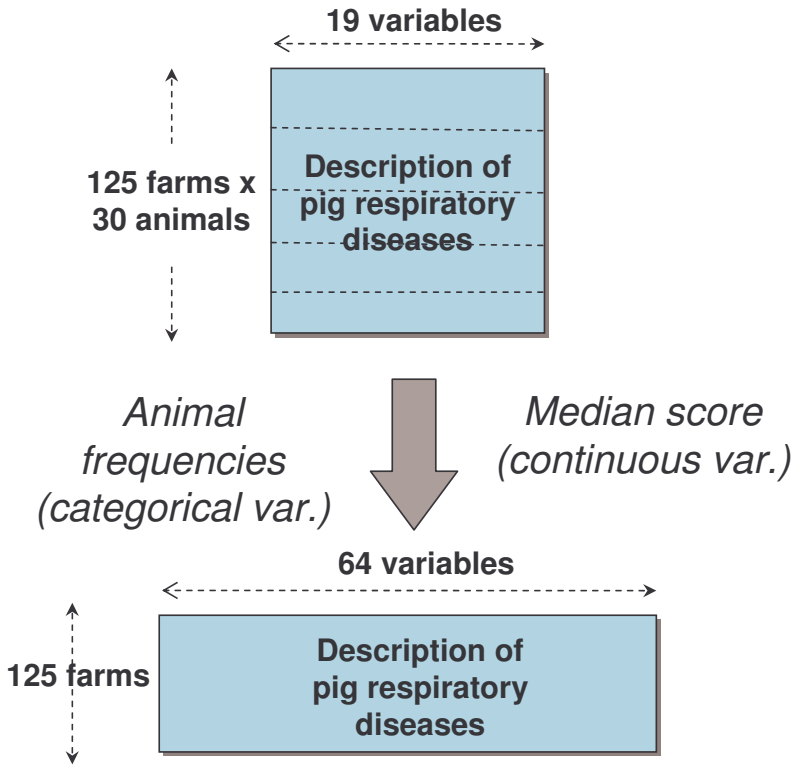
PB: FIND CORRELATIONS BETWEEN 3 CLASSICAL DATA TABLES OF DIFFERENT UNITS AND VARIABLES:
Table 1) Cracks description.
Table 2) Gap deviation of vertices of a grid at different periods compared to the initial model position.
Table 3) Gap depression from the ground.
ARE Transformed in ONE Symbolic Data Table where the concepts are interval of height

Hierarchical-Structured Data.



AFSSA: Study of pig respiratory diseases*

*C. Fablet, S. Bougeard (AFSSA)



Symbolic procedure

From numerical description of pigs to symbolic description of Farms

- Numerical variables

and

- Categorical variables

  are transformed in Bar Chart of the frequencies based on 30 animals,

Or in interval value variables

Step 1: Symbolic Description of Farms*

* SYROKKO Company eliezer@syrokko.com

- **Conclusion**

  In many cases COMPLEX DATA can be transformed in SYMBOLIC DATA.

OUTLINE

- What are Complex data?
- What are “symbolic data”?
- Symbolic Data are Complex data?
- Complex data are Symbolic Data after transformation ?
- **What is “Symbolic Data Analysis” (SDA)?**
- SDA gives a framework for Complex Data Analysis (CDA)?
- Open directions of research.

- **The Aim of SYMBOLIC DATA ANALYSIS?**

TO

EXTEND STATISTICS AND DATA MINING TO SYMBOLIC DATA TABLES DESCRIBING HIGHER LEVEL OBSERVATIONS (called "concepts") NEEDING VARIATION IN THEIR DESCRIPTION.

THE 4 MAIN TABLES IN SDA

1. **Standard data table T1:** observations x variables

Ex: players x variables each cell is numerical or categorical: Weight (Zidane) = 80

2. From T1 to **Random data table T2**: teams x variables

each cell is a Random Var. Ex: Weight (Lyon) = X_{21}

3. From T2 to **Symbolic data table T3** : teams x symb variables Ex: Weight (Lyon) = [80,95]

4. From T3 to **Random data table T4**: teams x variables

Example: an interval [80,95] which induces a random variable uniformly distributed on this interval.

RANDOM VARIABLES (T4) from SYMBOLIC DATA (T3)

•When the data are natively symbolic, which means that T1 and T2 are not known (ie only T3 is known) then the random variables X_{ij} of table T4 are defined from the Symbolic Data Table T3.

Examples:

Description of species of trees, mushroom, etc

Description of INSEE IRIS.

FROM SYMBOLIC DATA TABLE (T3) TO RANDOM VARIABLE DATA TABLE (T4)

Symbolic Data Table T3

	Y'_1		Y'_j	
C_1				
C_i				
C_k				

In each cell a symbolic data

Random Variable Data Table T4

	X_1		X_j	
C_1				
C_i			$\mathbf{X_{ij}}$	
C_k				

In each cell a Random Variable $C_i \longrightarrow D_j$

Symbolic Variables : C_i x Y'_j

(Classes of Observations) x (Random variables of Symbolic values)

$Y'j$ (Ci) = **Hij is a barchart value or an interval representing variation inside class C_i for the variable Y_j .**

Random Variables data table: C_i x X_j

(Classes of Observations) x (Random variables of random variable values)

Xj (Ci) = **Xij is a random variable:**

Xij (w) = **x** = a number or a category

From TABLE T4 with random variables in each cell To TABLE T5 with parameters in each cell

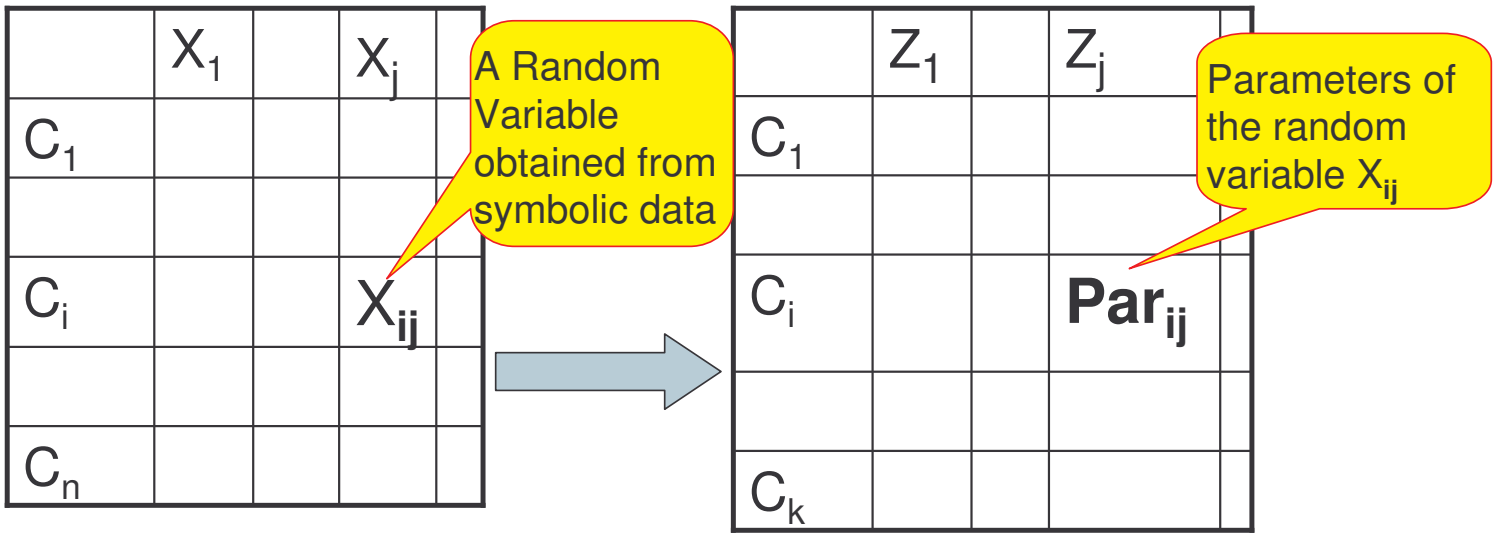


Table T4

Random data table T 4: w_i x X_j

(Classes of Observations) x (Rand Variables)
Xj (Ci) = **Xij is a random variable:**
Xij (w) = **x** = a number or a category

Table 5

Parametric data table 5: C_i x Z_j

(Classes of Observations) x (Random variables of parameter vector value)

Example: $\mathbf{Par_{ij}} = (\mu_{ij}, \sigma_{ij})$

From a random variable X_{ij} of T4 obtained from a Symbolic Data to its parameters of T5

Example: the case of interval symbolic data.

- The interval $[a_{ij} , b_{ij}]$ transformed in the random variable X_{ij} under the uniformity assumption has the following parameters:

➢ $\mu_{ij} = (a_{ij} + b_{ij})/2$

➢ $\sigma_{ij} = (a_{ij} - b_{ij})^{*2} /12$

(more details in Bock Diday (2000 Springer) in Bertrand Goupil chapter)

FOUR OPEN DIRECTION OF RESEARCH IN SDA

1. Non parametric:

Input: Symbolic data table 3

Output: Extending classical data mining to symbolic data table 3.

2. Semi parametric:

Input: Symbolic data table 3

Output: Copulas models from empirical distributions

3. Parametric:

Input: Table 4 of random variables induced from the symbolic data table 3 or Table 2 when the data are not natively symbolic.

Output: Table 5 of the parameters of the random variables of Table 2 or 4 .

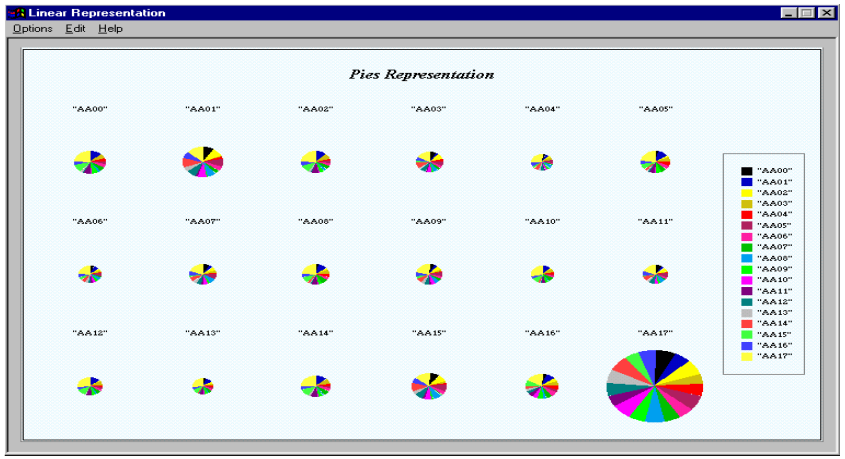
Descriptive statistic of Table 5 under models assumption (Gaussian, Dirichlet, Multinomial, etc.). Mixture decomposition.

4. Stochastic random data tables:

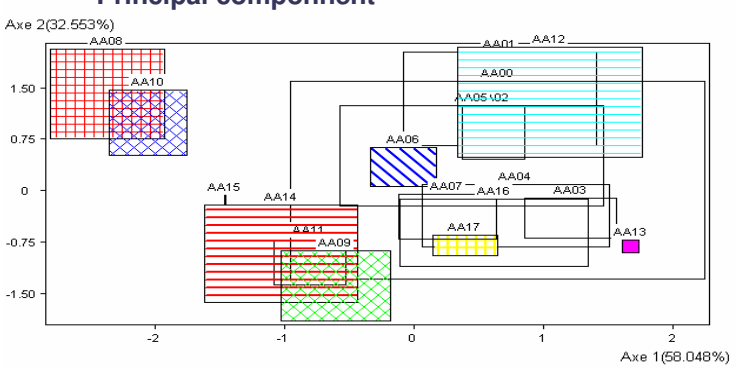
$X_{ij}^{(n)} \longrightarrow X_{ij}$ **for hudge data sets (data streams, cloud computers…) when the number of observations of the Data Table 1 increases does the classification structure (partition, hierarchy, pyramid, Galois Lattice,…) with their symbolic description converges?. (Only started: Stochastic Galois lattices with capacities).**

1. Non parametric: Extending Data Mining to Symbolic Data

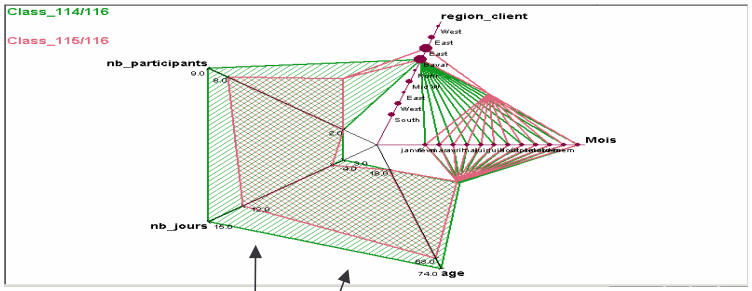
Kohonen map



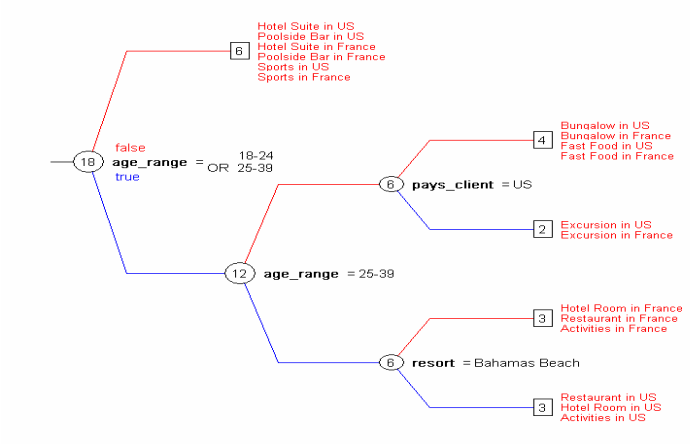
Principal componnent



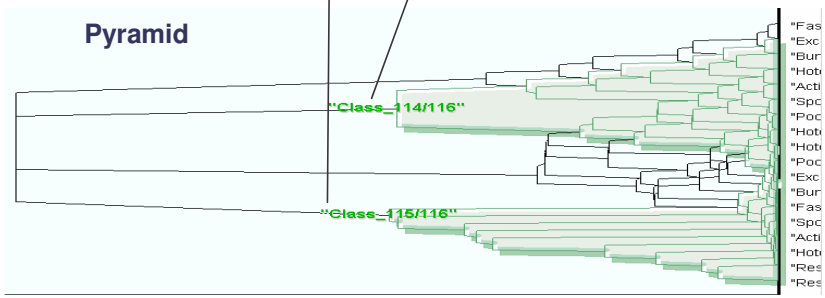
Zoom stars overlapping



Top down clustering tree or decision tree



Pyramid



1 Extending Data Mining to Symbolic Data: SOME RECENT ADVANCES

- PCA of bar chart data

- Symbolic Decision Trees, Regression

- Symbolic Text Mining

- Symbolic Time series

- Mixture Decomposition of symbolic Data by copulas

- Spatial Symbolic Clustering

FOUR OPEN DIRECTION OF RESEARCH IN SDA

1. Non parametric:

Input: Symbolic data table 3

Output: Extending classical data mining to symbolic data table 3.

2. Semi parametric:

Input: Symbolic data table 3

Output: Copulas models from empirical distributions

3. Parametric:

Input: Table 4 of random variables induced from the symbolic data table 3 or Table 2 when the data are not natively symbolic.

Output: Table 5 of the parameters of the random variables of Table 2 or 4 .

Descriptive statistic of Table 5 under models assumption (Gaussian, Dirichlet, Multinomial, etc.). Mixture decomposition.

4. Stochastic random data tables:

$X_{ij}^{(n)} \longrightarrow X_{ij}$ **for hudge data sets (data streams, cloud computers…) when the number of observations of the Data Table 1 increases does the classification structure (partition, hierarchy, pyramid, Galois Lattice,…) with their symbolic description converges?. (Only started: Stochastic Galois lattices with capacities).**

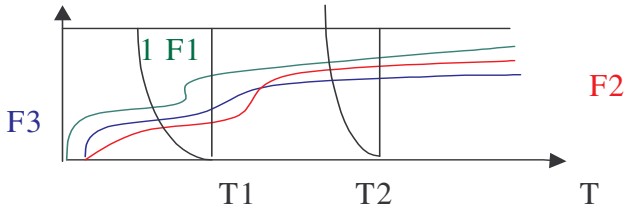
2. Semi parametric:

Input: Symbolic data table 3 where each cell contains a distribution.

Output: Copulas models from empirical distributions

DEFINITION OF A "POINT-DISTRIBUTION OF DISTRIBUTIONS"

$$G_{T_n}(x) = Pr(\{F_i \in F / F_i(T_n) \leq x\}) = card(\{F_i \in F / F_i(T_n) \leq x\})/N$$



DEFINITION OF A "K-POINT JOINT DISTRIBUTION OF DISTRIBUTIONS"

$$H_{T1,\ldots,T_k}(x_1,\ldots,x_k) = Pr(\{F_i \in F / F_i(T_1) \leq x_1\} \wedge \ldots \wedge \{F_i \in F / F_i(T_k) \leq x_k\}).$$

PROPOSITION 1

. G_{T_n} IS A DISTRIBUTION.

. $H_{T1,\ldots,T_k}$ IS A K-DIMENSIONAL JOINT DISTRIBUTION FUNCTION WITH MARGIN

$G_{T1},\ldots, G_{Tk}$

WHAT LINK BETWEEN

. THE JOINT H

. THE MARGINAL G_{Ti} **?**

$H_{T1,\ldots,Tk}(X_1,\ldots,X_k) = \mathbf{C}(G_{T1}(X_1),\ldots. G_{Tk}(X_k))$

C IS A K-COPULAS

. DEFINITION OF C ?

. EXISTENCE ?

. PROPERTIES ? (UNICITY,…)

(Vrac, Cuvelier Dissertation…)

FOUR OPEN DIRECTION OF RESEARCH IN SDA

1. Non parametric:

Input: Symbolic data table 3

Output: Extending classical data mining to symbolic data table 3.

2. Semi parametric:

Input: Symbolic data table 3

Output: Copulas models from empirical distributions

3. Parametric:

Input: Table 4 of random variables induced from the symbolic data table 3 or from Table 2 when the data are not natively symbolic.

Output: Table 5 of the parameters of the random variables of Table 2 or 4 .

Descriptive statistic of Table 5 under models assumption (Gaussian, Dirichlet, Multinomial, etc.). Mixture decomposition, Biclassification, Partitioning, Galois Lattice, Pyramid , hierarchy

4. Stochastic random data tables:

$X_{ij}^{(n)} \longrightarrow X_{ij}$ **for hudge data sets (data streams, cloud computers…) when the number of observations of the Data Table 1 increases does the classification structure (partition, hierarchy, pyramid, Galois Lattice,…) with their symbolic description converges?. (Only started: Stochastic Galois lattices with capacities).**

3. Parametric: Statistical description of Table 5

RECALL:

Table 1 Players x standard Random Variables (RV), Table 2 Teams x (RV of RV),

Table 3 Teams x Symbolic Data (SD) Example: Lyon Weight [80,95],

Table 4 Teams x RV Example: Uniform RV associated to the interval [80,95],

Table 5 Each cell contains the parameters of the T4 Random Variables,

Example: mean = (80 + 95)/ 2 , mean square = $(95 - 80)^2 / 12$

EXAMPLE of descriptive statistic of Table 5:

Sample Mean, Sample Variance (Bertrand and Goupil (2000))

For an interval-valued random variable Y, the symbolic sample mean is given by

$$\bar{Y} = \frac{1}{2m} \sum_{u \in E} (b_u + a_u)$$

and the symbolic sample variance is given by

$$S^2 = \frac{1}{3m} \sum_{u \in E} (b_u^2 + b_u a_u + a_u^2) - \frac{1}{4m^2} [\sum_{u \in E} (b_u + a_u)]^2,$$

for observations $Y_u = [a_u, b_u), u = 1, \ldots, m$

INFERENCE ON THE PARAMETRIC DATA TABLE 5

Sample mean, variance, correlation of parametric data table 5 follow which model?

- Infering Models on mean, mean square, correlation ,
- Finding models on classes:
  - Mixture decomposition,
  - Biclassification…
  - Lattices, Pyramids, hierarchies...

CONCLUSION

- Symbolic Data (SD) are complex data as they cannot be transformed in standard data.
- Complex data can be transformed in SD.
- Therefore SDA is a tool for extracting knowledge from classes of standard data or from complex data.
- Much open research for non parametric or parametric SDA from the symbolic data or from their induced random variable distributions.
- Much to be done for stochastic symbolic data for cloud computers

THREE SDA Books

WILEY, 2008
"Symbolic Data Analysis and the SODAS software." 457 pages
E. Diday, M. Noirhomme , (www.wiley.com)

WILEY, 2006
L. Billard , E. Diday "Symbolic Data Analysis, conceptual statistic and Data Mining".www.wiley.com

SPRINGER, 2000 :
"Analysis of Symbolic Data"
H.H., Bock, E. Diday, Editors . 450 pages.

Références

Afonso F., Billard L., E. Diday (2004) : Régression linéaire symbolique avec variables taxonomiques, Revue RNTI, Extraction et Gestion des Connaissances (EGC 2004),G.Hébrail et al. Eds, Vol. 1, p. 205-210, Cépadues, 2004.

- Afonso F., Diday E. (2005) : Extension de l'algorithme Apriori et des règles d'association aux cas des données symboliques diagrammes et intervalles, Revue RNTI, Extraction et Gestion des Connaissances (EGC 2005), Vol. 1, pp 205-210, Cépadues, 2005.
- Aristotle (IV BC): Organon Vol. I Catégories, II De l'interprétation. J. Vrin edit. (Paris) (1994).
- Arnault A., Nicole P. (1662) : La logique ou l'art de penser, Froman, Stutgart (1965).
- Appice A., D'Amato C., Esposito F., Malerba D. (2006): Classification of Symbolic Objects: A Lazy Learning Approach. Intelligent Data Analysis, 10 (4), 301 – 324
- .Bezerra B. L. D., De Carvalho F.A.T. (2004): A symbolic approach for content-based information filtering. Information Processing Letters, 92 (1), 45-52.
- Billard L. (2004): Dependencies in bivariate interval-valued symbolic data.. In: Classification, Clustering and New Data Problems . Proc. IFCS'2004. Chicago. Ed. D. Banks. Springer Verlag, 319-354.
- Billard L., Diday E. (2006): Symbolic Data Analysis: Conceptual Statistics and Data Mining. To be published by Wiley.

Billard L., Diday E. (2005): Histograms in symbolic data analysis 2005. Intern Stat. Inst. 55.
Bravo Llatas M.C. (2004): Análisis de Segmentación en el Análisis de Datos Simbólicos. Ed. Universidad Complutense de Madrid. Servicio de Publicaciones. ISBN:8466917918. (http://www.ucm.es/BUCM/tesis/mat/ucm-t25329.pdf)
Brito, P. (2005) : Polaillon, G., Structuring Probabilistic Data by Galois Mathématiques et Sciences Humaines, 43ème année, nº 169, (1), pp. 77-104.
Brito, P. (2002): Hierarchical and Pyramidal Clustering for Symbolic Data, Journal of the Japanese Society of Computational Statistics, Vol. 15, Number 2, pp. 231-244.
Caruso C., Malerba D., Papagni D. (2005). Learning the daily model of network traffic. In M.S. Hacid, N.V. Murray, Z.W. Ras, S. Tsumoto (Eds.) Foundations of Intelligent Systems, 15th International Symposium, ISMIS'2005, Lecture Notes in Artificial Intelligence, 3488, 131-141, Springer, Berlin,
Germania. Cazes, P., Chouakria, A., Diday, E. Schektman, Y. (1997) Extension de l'analyse en composantes principales à des données de type intervalle, Revue de Statistique Appliquée XIV(3), 5–24.
Ciampi A., Diday E., Lebbe J., Perinel E., R. Vignes (2000): Growing a tree classifier with imprecise data. Pattern. Recognition letters 21, pp 787-803.

- De Carvalho F.A.T., Eufrasio de A. Lima Neto, Camilo P.Tenerio (2004): A new method to fit a linear regression model for interval-valued data. In: Advances in Artificial Intelligence: Proceedings of the Twenty Seventh German Conference on Articial Intelligence (eds. S. Biundo, T. Fruchrirth, and G. Palm). Springer-Verlag, Berlin, 295-306.De Carvalho F.A.T., De Souza R., Chavent M., Y. Lechevallier (2006): Adaptive Hausdorff distances and dynamic clustering of symbolic interval data. Pattern Recognition Letters, 27 (3), 167-179
- De Carvalho F.A.T., Brito P., Bock H. H. (2006), Dynamic Clustering for Interval Data Based on L_2 Distance, Computational Statistics, accepted for publication.
- De Carvalho, F. A. T. (1995): Histograms In Symbolic Data Analysis. Annals of Operations Research,Volume 55, Issue 2, 229-322.
- De Souza, R. M. C. R. and De Carvalho, F. A. T. (2004): Clustering of Interval Data based on City-Block Distances. Pattern Recognition Letters, Volume 25, Issue 3, 353-365.
- Diday E. (1987 a): The symbolic aproach in clustering and related methods of Data Analysis. In "Classification and Related Methods of Data Analysis", Proc. IFCS, Aachen, Germany. H. Bock ed.North-Holland.
- Diday E. (1987 b): Introduction à l'approche symbolique en Analyse des Données. Première Journées Symbolique-Numérique. Université Paris IX Dauphine. Décembre 1987.

- Diday E. (1989): Introduction à l'Analyse des Données Symboliques. Rapport de Recherche INRIA N° 1074 (August 1989). INRIA Rocquencourt 78150. France.
- Diday E. (1991) : Des objets de l'Analyse des Données à ceux de l'Analyse des Connaissances. In « Induction Symbolique et Numérique à partir de données ». Y. Kodratoff, Diday E. Editors. CEPADUES-EDITION.ISBN 2.85428.282 5.
- Diday E. (2000): L'Analyse des Données Symboliques : un cadre théorique et des outils pour le Data Mining. In : E. Diday, Y. Kodratoff, P. Brito, M. Moulet "Induction symbolique numérique à partir de données". Cépadues. 31100 Toulouse. www.editions-cepadues.fr. 442 pages.
- Diday E. (2002): An introduction to Symbolic Data Analysis and the Sodas software. Journal of Symbolic Data Analysis. Vol. 1, n° 1. International Electronic Journal. www.jsda.unina2.it/JSDA.htm.
- Diday E., Esposito F. (2003): An introduction to Symbolic Data Analysis and the Sodas Software IDA. International Journal on Intelligent Data Analysis". Volume 7, issue 6. (Decembre).
- Diday E., Emilion R. (2003): Maximal and stochastic Galois Lattices. Journal of Discrete Applied Mathematics, Vol. 127, pp. 271-284.
- Diday E. (2004): Spatial Pyramidal Clustering Based on a Tessellation. Proceedings IFCS'2004, In Banks andal. (Eds.): Data Analysis, Classification and Clustering Methods

- Diday E., Vrac M. (2005): Mixture decomposition of distributions by Copulas in the symbolic data analysis framework. Discrete Applied Mathematics (DAM). Volume 147, Issue 1, 1 April, Pages 27-41.
- E. Diday (2005): Categorization in Symbolic Data Analysis. In handbook of categorization in cognitive science. Edited by H. Cohen and C. Lefebvre. Elsevier editor. http://books.elsevier.com/elsevier/?isbn=0080446124
- Diday E.(1995): Probabilist, possibilist and belief objects for knowledge analysis. Annals of Operations Research. 55, pp. 227-276.
- Diday E., Murty N. (2005): Symbolic Data Clustering. In Encyclopedia of Data Warehousing and Mining . John Wong editor . Idea Group Reference Publisher.
- Duarte Silva, A. P., Brito, P. (2006): Linear Discriminant Analysis for Interval Data,Computational Statistics, accepted for publication.
- Gioia, F. and Lauro, N.C. (2005) Basic Statistical Methods for Interval Data, Statistica applicata, 1.
- Gioia, F. and Lauro, N.C. (2006): Principal Component Analysis on Interval Data, Computational statistics, In press.
- Hardy, A. and Lallemand, P. (2002): Determination of the number of clusters for symbolic objects described by interval variables, In Studies in Classification, Data Analysis, and Knowledge Organization, Proceedings of the IFCS'02 Conference, 311-318.

- Hardy, A, Lallemand, P. and Lechevallier, Y. (2002) : La détermination du nombre de classes pour la méthode de classification symbolique SCLUST, Actes des Huitièmes Rencontres de la Société Francophone de Classification, 27-31
- Hardy, A. and Lallemand, P. (2004): Clustering of symbolic objects described by multi-valued and modal variables, In Studies in Classification, Data Analysis, and Knowledge Organization, Proceedings of the IFCS'04 Conference, 325-332
- Hardy, A. (2004): Les méthodes de classification et de determination du nombre de classes: du classique au symbolique, In M. Chavent, O. Dordan, C. Lacomblez, M. Langlais, B. Patouille (Eds), Comptes rendus des Onzièmes Rencontres de la Société Francophone de Classification, 48-55
- Hardy, A. (2005): Validation in unsupervised symbolic classification, Proceedings of the Meeting "Applied Stochastic Models and Data Analysis " (ASMDA 2005), 379-386
- Irpino, A. (2006): Spaghetti PCA analysis: An extension of principal components analysis to time dependent interval data. Pattern Recognition Letters, Volume 27, Issue 5, 504-513.
- Irpino, A., Verde, R. and Lauro N. C. (2003): Visualizing symbolic data by closed shapes, Between Data Science and Applied Data Analysis, Shader-Gaul-Vichi eds., Springer, Berlin, pp. 244-251.

- Lauro, N.C., Verde, R. and Palumbo, F. (2000): Factorial Data Analysis on Symbolic Objects under cohesion constrains In: Data Analysis, Classification and related methods, Springer-Verlag, Heidelberg
- M. Limam, E. Diday, S. Winsberg (2004): Symbolic Class Description with Interval Data. Journal of Symbolic Data Analysis, 2004, Vol 1
- D. Malerba, F. Esposito, M. Monopoli (2002): Comparing dissimilarity measures for probabilistic symbolic objects.In A. Zanasi, C. A. Brebbia, N.F.F. Ebecken, P. Melli (Eds.)Data Mining III, Series Management Information Systems, Vol 6, 31-40, WIT Press, Southampton, UK. Mballo C., Asseraf M., E. Diday (2004): Binary tree for interval and taxonomic variables. A Statistical Journal for Graduates Students"Volume 5, Number 1, April 2004.
- Milligan , G.W., Cooper M.C. (1985): An examination of procedures for determining the number of clusters in a data set. Psychometrica 50, 159-179.
- MenesesE., Rodríguez-Rojas O. (2006): Using symbolic objects to cluster web documents. WWW 2006: 967-968.

- Noirhomme-Fraiture, M. (2002): Visualization of Large Data Sets : the Zoom Star Solution, Journal of Symbolic Data Analysis, vol. 1, July.
- <http://www.jsda.unina2.it/>http://www.jsda.unina2.it
- Prudêncio R. B. C., Ludermir T., F. de A. T. De Carvalho (2004): A Modal Symbolic Classifier for selecting time series models.Pattern Recognition Letters, 25 (8), 911-921.
- Rodriguez O. (2000): "Classification et modèles linéaires en Analyse des Données Symboliques". Thèse de doctorat, University Paris 9 Dauphine.
- Schweizer B. (1985) "Distributions are the numbers of the futur" . Proc. sec. Napoli Meeting on "The mathematics of fuzzy systems". Instituto di Mathematica delle Faculta di Mathematica delle Faculta di Achitectura, Universita degli studi di Napoli. p. 137-149.
- Schweizer B. , Sklar A. (2005): Probabilist metric spaces . Dover Publications INC. Mineola, New-York.Soule A., K. Salamatian, N. Taft, R. Emilion (2004): “Flow classfication by histograms” ACM SIGMETRICS, New York. http://rp.lip6.fr/~soule/SiteWeb/Publication.php
- Stéphan V. (1998): "Construction d'objets symboliques par synthèse des résultats de requêtes". (1998). Thesis. Paris IX Dauphine University.
- Vrac M, Diday E., Chédin A. (2004) : Décomposition de mélange de distributions et application à des données climatiques. Revue de Statistique Appliquée, 2004, LII (1), 67-96.
- Vrac M, Diday E., Chédin A. (2004) : Décomposition de mélange de distributions et application à des données climatiques.Revue de Statistique Appliquée, 2004, LII (1), 67-96.