NEW ADVANCES IN SYMBOLIC DATA ANALYSIS and SPATIAL CLASSIFICATION.

Edwin Diday
Paris Dauphine University

Remembering

Suzanne WINSBERG

This talk is dedicated to her…

OUTLINE

PART 1: SYMBOLIC DATA ANALYSIS

- The two levels of statistical units: individuals, concepts
- What are Symbolic Data?
- What is Symbolic data analysis?
- Why and when Symbolic Data Analysis?
- Future of SDA

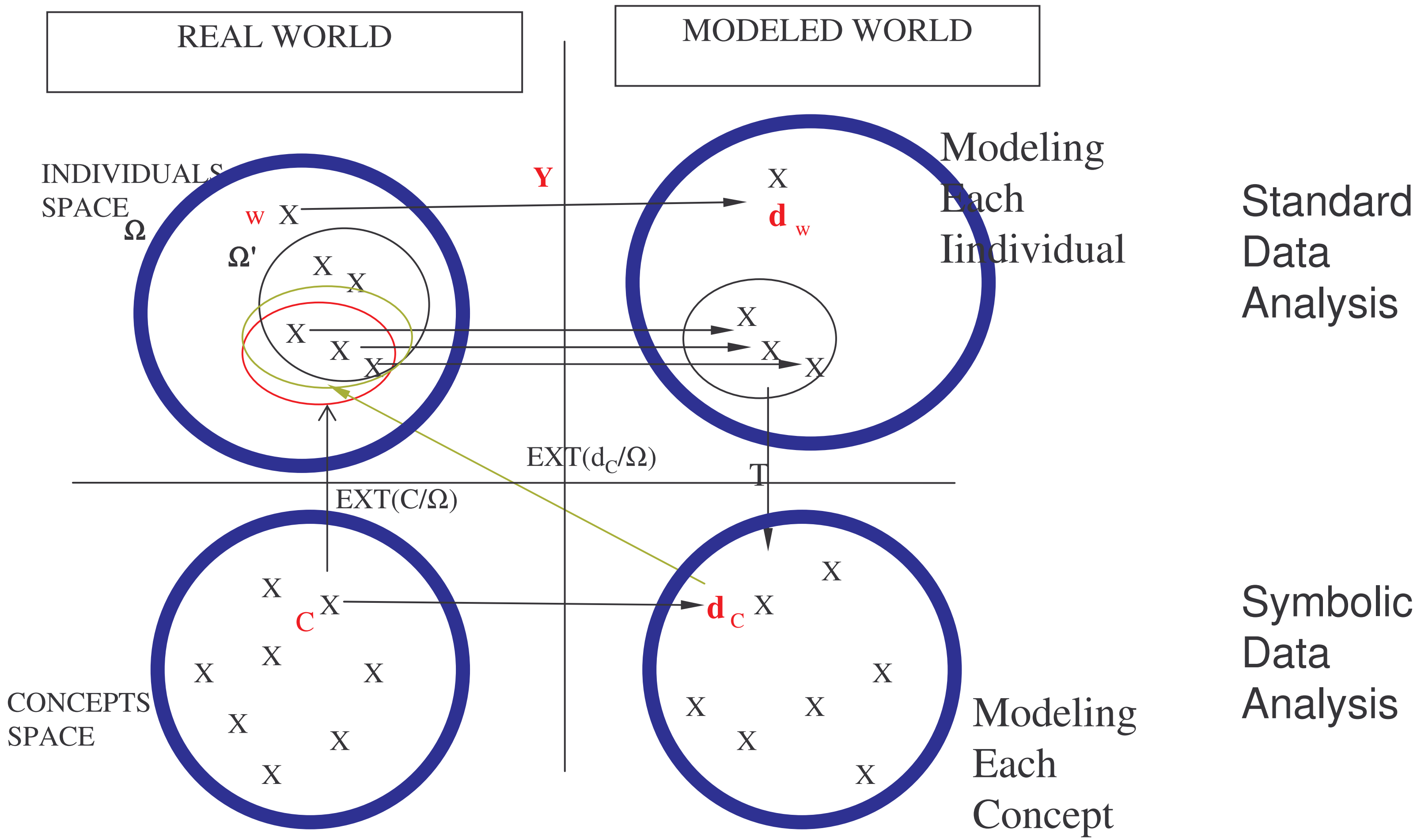
PART 2: SPATIAL CLASSIFICATION

Symbolic Data Analysis software:

SODAS and SYR

THE TWO LEVELS OF STATISTICAL UNITS:

- INDIVIDUALS
- CONCEPTS



REAL WORLD
MODELED WORLD
INDIVIDUALS SPACE
Ω
Ω'
w
Y
d w
Modeling Each Iindividual
Standard Data Analysis
EXT(d_C/Ω)
T
EXT(C/Ω)
C
d C
CONCEPTS SPACE
Modeling Each Concept
Symbolic Data Analysis

BASIC IDEAS OF SDA

- **TWO LEVELS OF OBJECTS:**
  - **First level: Individuals**
  - **Second level: categories, classes or concepts (intent,extent)**
- **SECOND LEVEL UNITS CAN BE CONSIDERED AS NEW STATISTICAL UNITS.**
- **A CONCEPT IS DESCRIBED BY THE VARIATION OF THE CLASS OF INDIVIDUALS THAT IT REPRESENTS:**
- **THIS PRODUCES SYMBOLIC DATA.**

FROM INDIVIDUALS TO CONCEPTS

Classical : individuals

Birds

Inhabitant

Players (Zidane,…)

Image

Sold clothes

Trace of WEB Usage

Patients after heart attack

Mobile users

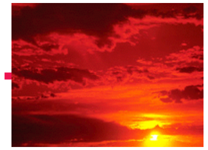
Symbolic : concepts

Species of birds

Regions

Team (Marseille, …)

Type of image (sunset,,..

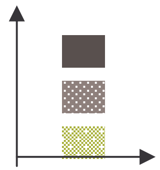


Shops

Users

Trajectory of patients in hospitals

Consuming level



- **WHAT ARE SYMBOLIC DATA?**

SYMBOLIC DATA

TEAM OF THE FRENCH CUP	WEIGHT	NATIONALITY	NB OF GOALS
DIJON	[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.

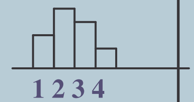
Therefore each cell can contain:

A number, an interval, a sequence of categorical values, a sequence of weighted values as a histograms, 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.

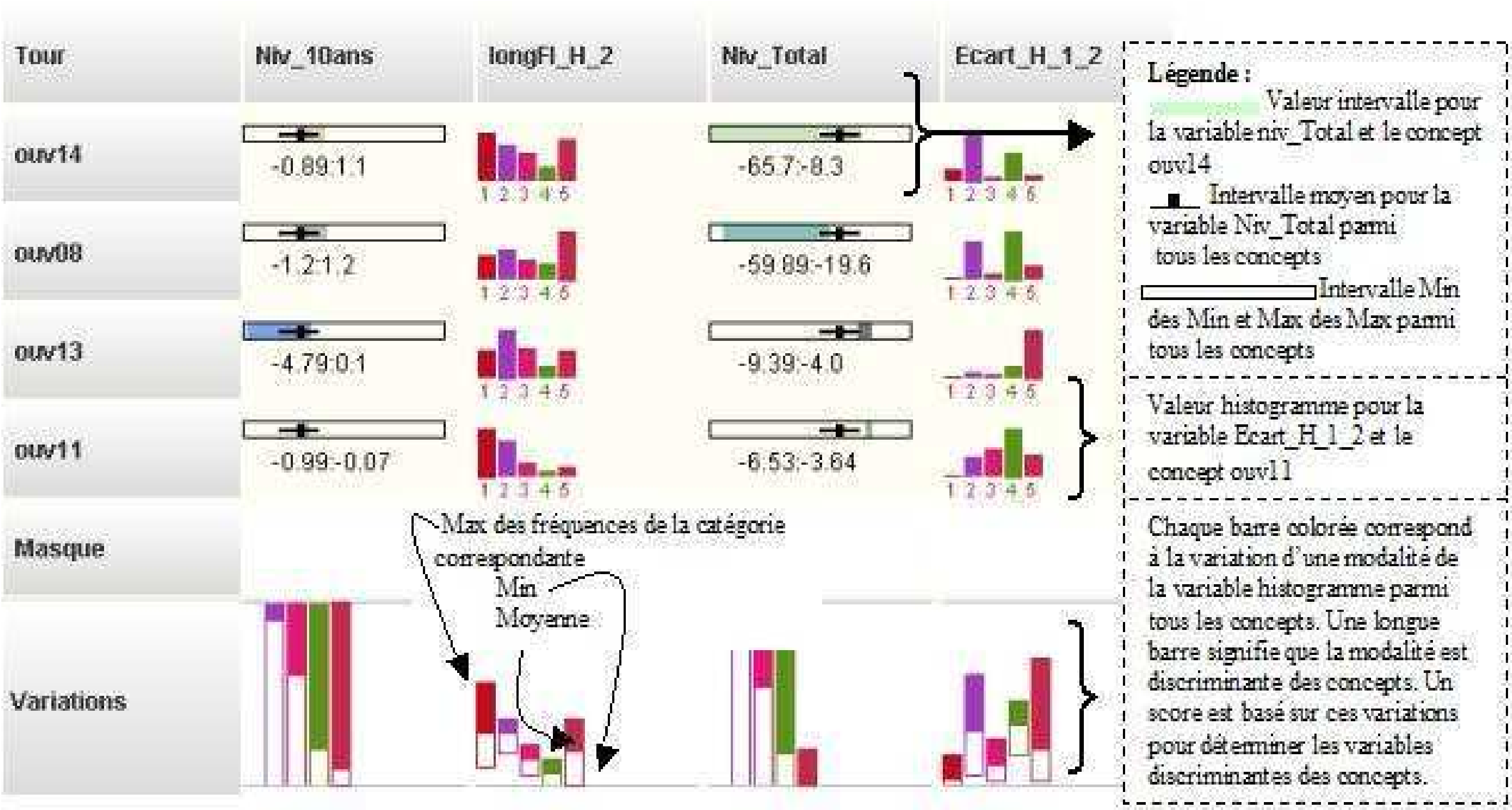
How to conserve correlation and explain it?

patients	Region	Cardiology Expenses	Dentistry Expenses	Town	Insurance
i1	R1	12.5	3,5	Lyon1	Type 3
i2	R1	9.6	2,1	Paris3	Type 2
i3	R1	11.4	6.5	Lyon1	Type 4
i4	R2	3.2	1,6	Paris1	Type 1
i5	R2	7.1	4,8	Lyon2	Type 2

Concept	Card. Expenses	Dentistry Exp.	Town	Insurance	Cor(card, dentist)
R1	[9.6, 12.5]	[2.1, 6.5]	{Lyon1, Paris 3}		Cor_{R1} (cardi, dent)
R2	[3.2, 7.1]	[1.6, 4.8]	Paris 3		Cor_{R2} (cardi, dent)
R3	[9.2, 10.1]	[6.2, 8.1]	Pau 1		Cor_{R3} (cardi, dent)
R4	[5, 8.4]	[7.3, 9.4]	Pau 4		Cor_{R4} (cardi, dent)

Then, a symbolic regression or symbolic decision tree can explain the correlation.

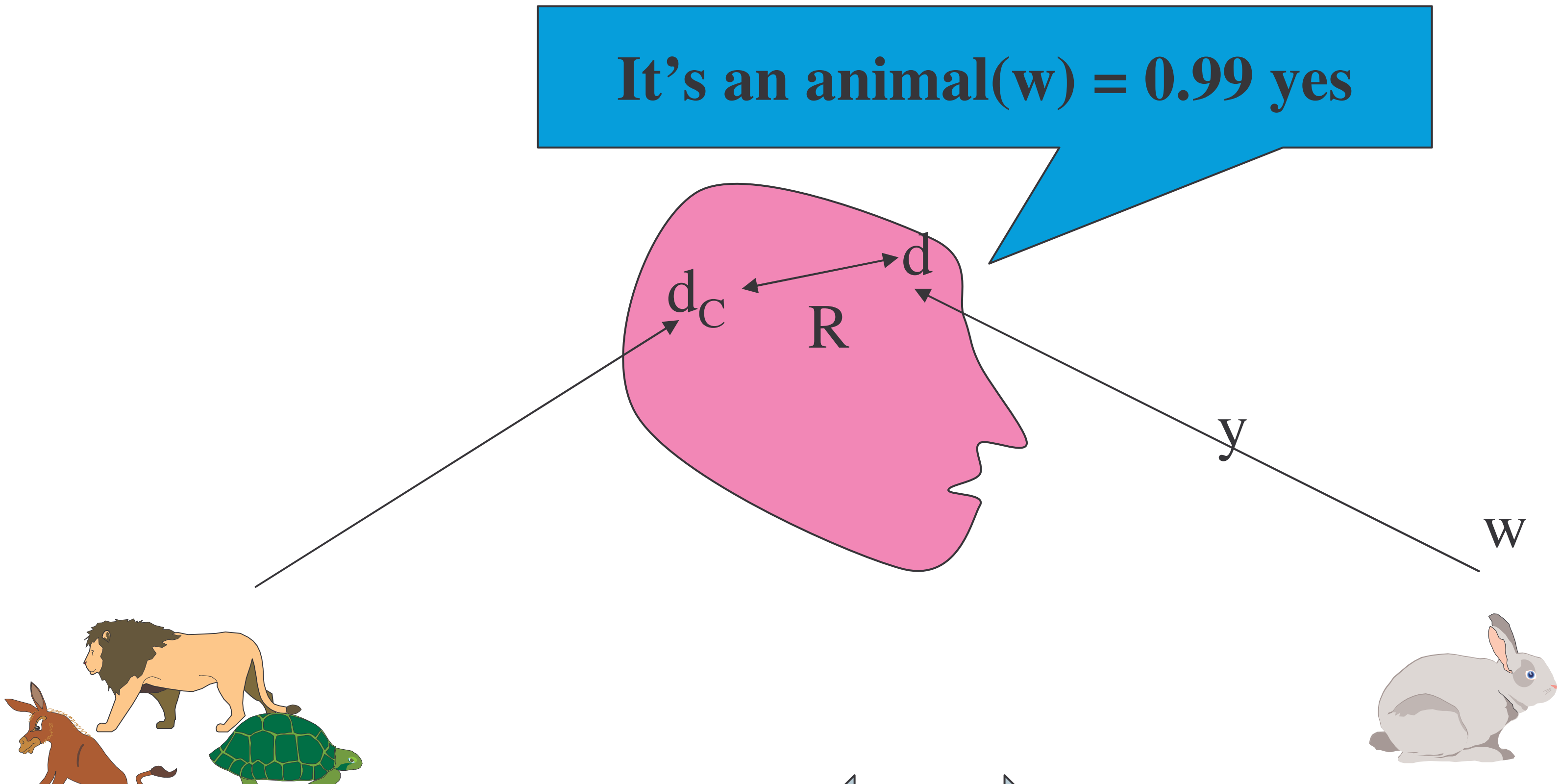
Symbolic Data Table
(SYR software)



How to model concepts?

- By the so called “Symbolic Objects”

SYMBOLIC OBJECT



$$\mathbf{S = (a, R, d_C)} \Leftrightarrow \mathbf{a(w) = [y(w)Rd_C]}$$

TWO KINDS OF SYMBOLIC OBJECTS

BOOLEAN SYMBOLIC OBJECTS

S = (a, R, d1)

d1 = $\{12, 20, 28\} \times \{\text{employee, worker}\}]$

R = $(\subseteq, \subseteq)$,

a(w) = $[\text{age}(w) \subseteq \{12, 20, 28\}] \wedge [\text{SPC}(w) \subseteq \{\text{employee, worker}\}]$

a(w) $\in$ **{TRUE, FALSE}.**

THE MEMBERSHIP FUNCTION« a » MODAL CASE

S = (a, R, d):

a(w) = [age(w) R_1 {(0.2)12, (0.8) [20 ,28]}] $\wedge$ [SPC(w) R_2 {(0.4)employee, (0.6)worker}]

a(w) $\in$ [0,1].

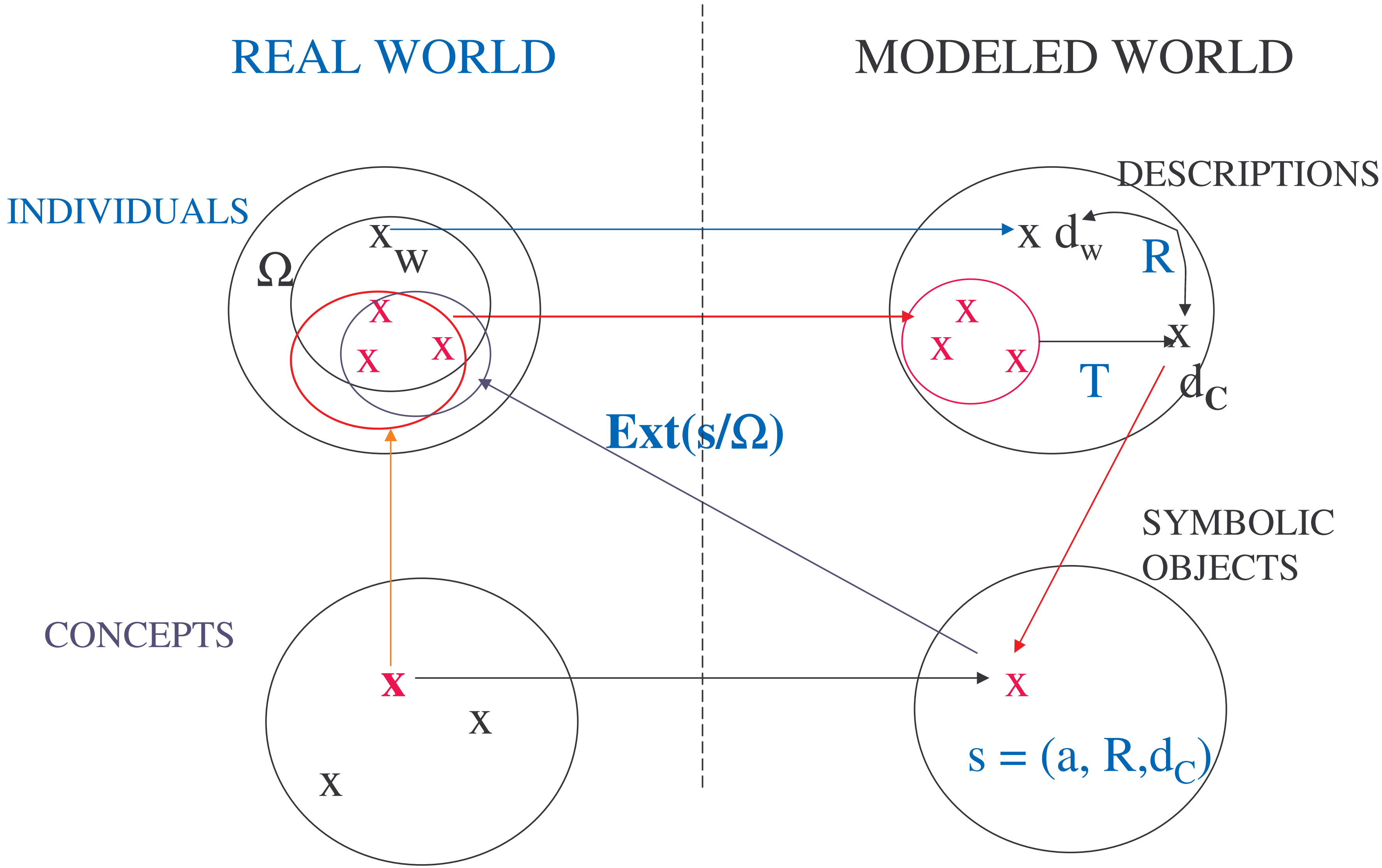
First approach: simple or flexible matching

R= (R_1, R_2): $r \, R_i \, q = \sum_{j=1,k} r_j \, q_j \, e^{(r_j - \min(r_j, q_j))}$.

Second approach:

Probabilistic: if dependencies, copulas,

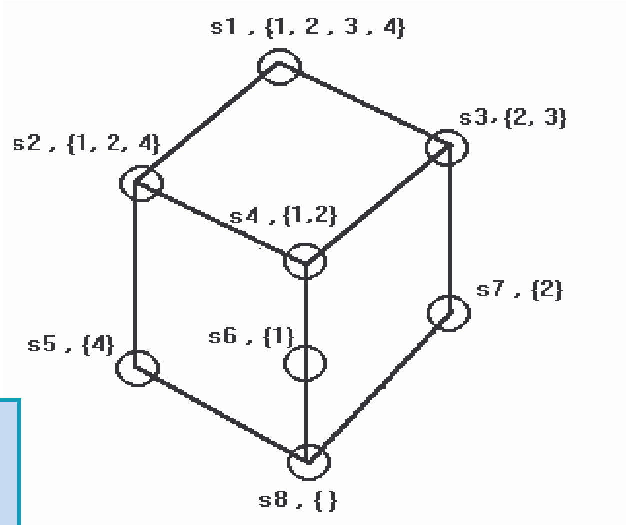
QUALITY CONTROL CONFIRMATORY SDA



THE SYMBOLIC DATA TABLE

	Y1	Y2	Y3
W1	{a, b}	∅	{g}
W2	∅	∅	{g, h}
W3	{c}	{e, f}	{g, h, i}
W4	{a, b, c}	{e}	{h}

Lattice obtained from the symbolic Data Table



Symbolic objects obtained From the Symbolic Lattice.

$s_2 : a_2(w) = [y_2(w) \subseteq \{e\}] \wedge [y_3(w) \subseteq \{g,h\}]$,

$\mathrm{Ext}(s_2) = \{1, 2, 4\}$

$s_3 : a_3(w) = [y_1(w) \subseteq \{c\}]$,

$\mathrm{Ext}(s_3) = \{2, 3\}$

$s_4 : a_4(w) = [y_1(w) \subseteq \{a,b\}] \wedge [y_2(w) = \varnothing] \wedge [y_3(w) \subseteq \{g,h\}]$,

$\mathrm{Ext}(s_4) = \{1, 2\}$

- **WHAT IS SYMBOLIC DATA ANALYSIS?**

TO

EXTEND STATISTICS AND DATA MINING TO SYMBOLIC DATA TABLES DESCRIBING HIGHER LEVEL UNITS NEEDING VARIATION IN THEIR DESCRIPTION.

SYMBOLIC DATA ANALYSIS TOOLS HAVE BEEN DEVELOPPED

- Graphical visualisation of Symbolic Data
- Correlation, Mean, Mean Square Histogram of a symbolic variable
- Dissimilarities between symbolic descriptions
- Clustering of symbolic descriptions
- S-Kohonen Mappings
- S-Decision Trees
- S-Principal Component Analysis
- S-Discriminant Factorial Analysis
- S-Regression
- Etc...

Why Symbolic Data Analysis?

1) From standard statistical units to concepts,
the statistic is not the same!

2) Symbolic Data cannot be reduced to classical data!

From standard statistical units to concepts,
The statistic is not the same!

On an island : Three species of 600 birds together: 400 swallows, 100 ostriches, 100 penguins.

Bird	Species	Flying	Size (cm)
1	penguins	No	80
2	swallows	yes	30
600	ostriches	No	125

Symbolic Data Table

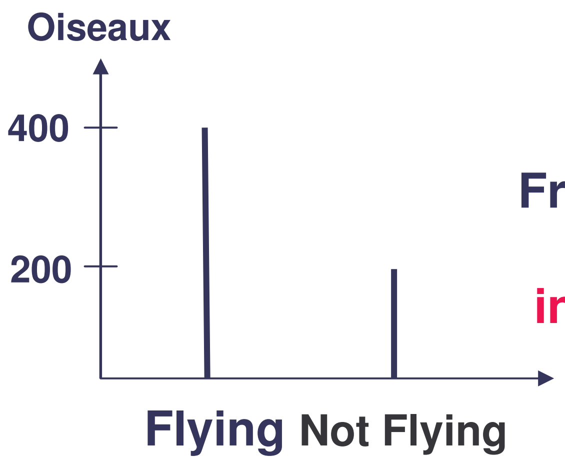
Species	Fly	Couleur	Taille	Migr
swallows	yes	0.3b,0.7grey	[25, 35]	Yes
ostriche	No	0.1black,0.9g	[85,160]	No
Penguin	No	0.5b,0.5grey	[70, 95]	Yes

The species are the new units

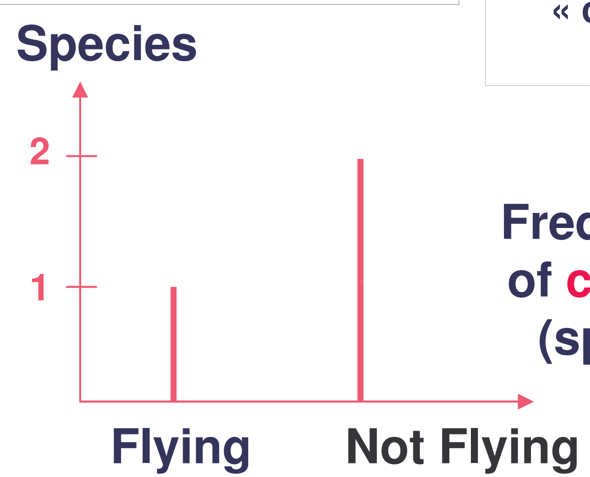
swallows, ostriches, and penguins are the "concepts"

The variation due to the individuals of each species produces symbolic data

« Migration » is an added variable at the the « concepts » level.



Frequencies of individuals



Frequencies of concepts (species)

WHY SYMBOLIC DATA CANNOT BE REDUCED TO CLASSICAL DATA?

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	Poids Min	Poids Max	Taille Min	Taille Max	Eur	Afr
Very good	80	95	1.70	1.95	0. 7	0.3

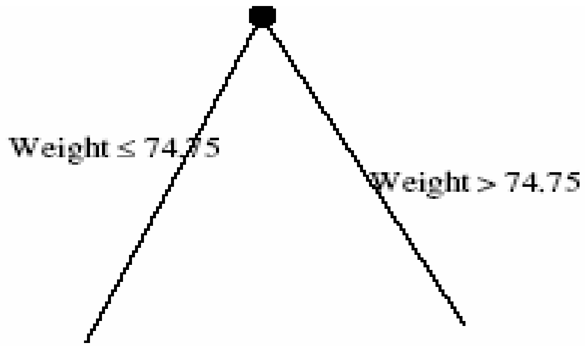
Concern:

The initial variables are lost and the variation is lost!

Divisive Clustering   or Decision tree

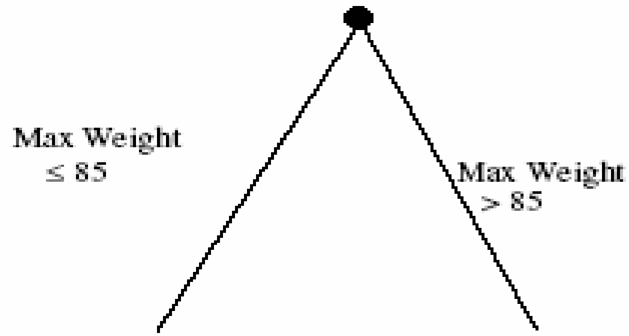
Symbolic Analysis

Weight



Classical Analysis

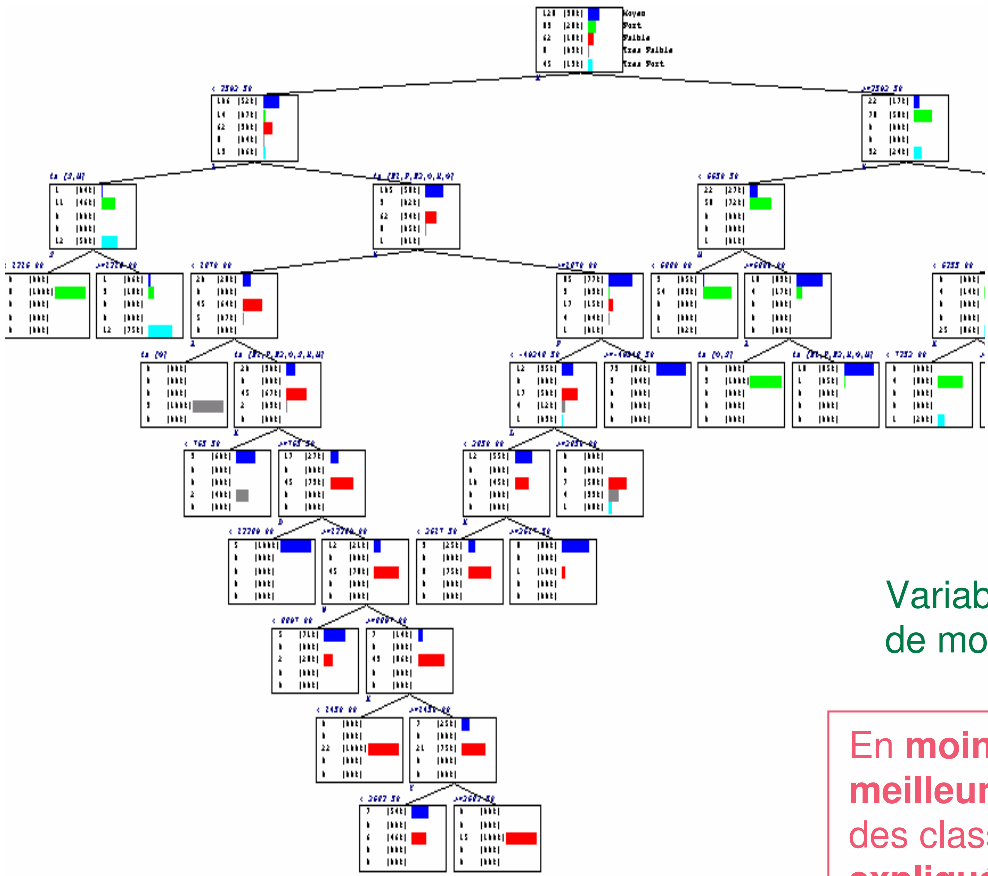
Max Weight



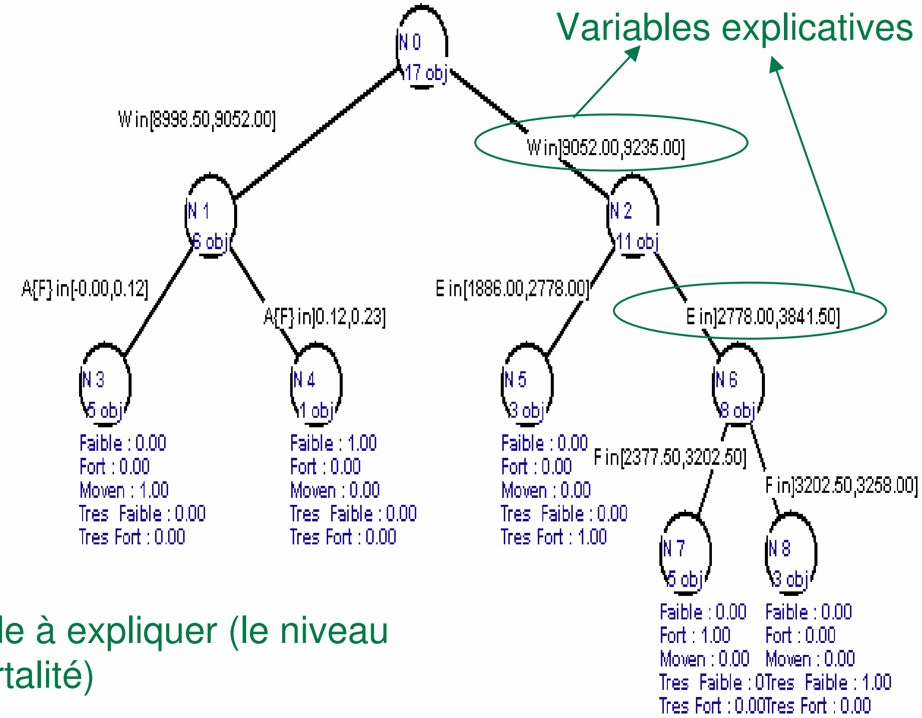
Classique / symbolique : une comparaison

Arbres de décision établis sur 1000 données initiales (patients) que l'on veut regrouper en classes homogènes suivant une même trajectoire d'hospitalisation. Variable à expliquer (ex. la mortalité) et des variables explicatives cliniques-biologiques.

Arbre « classique » sur les patients



Arbre « symbolique » sur les trajectoires

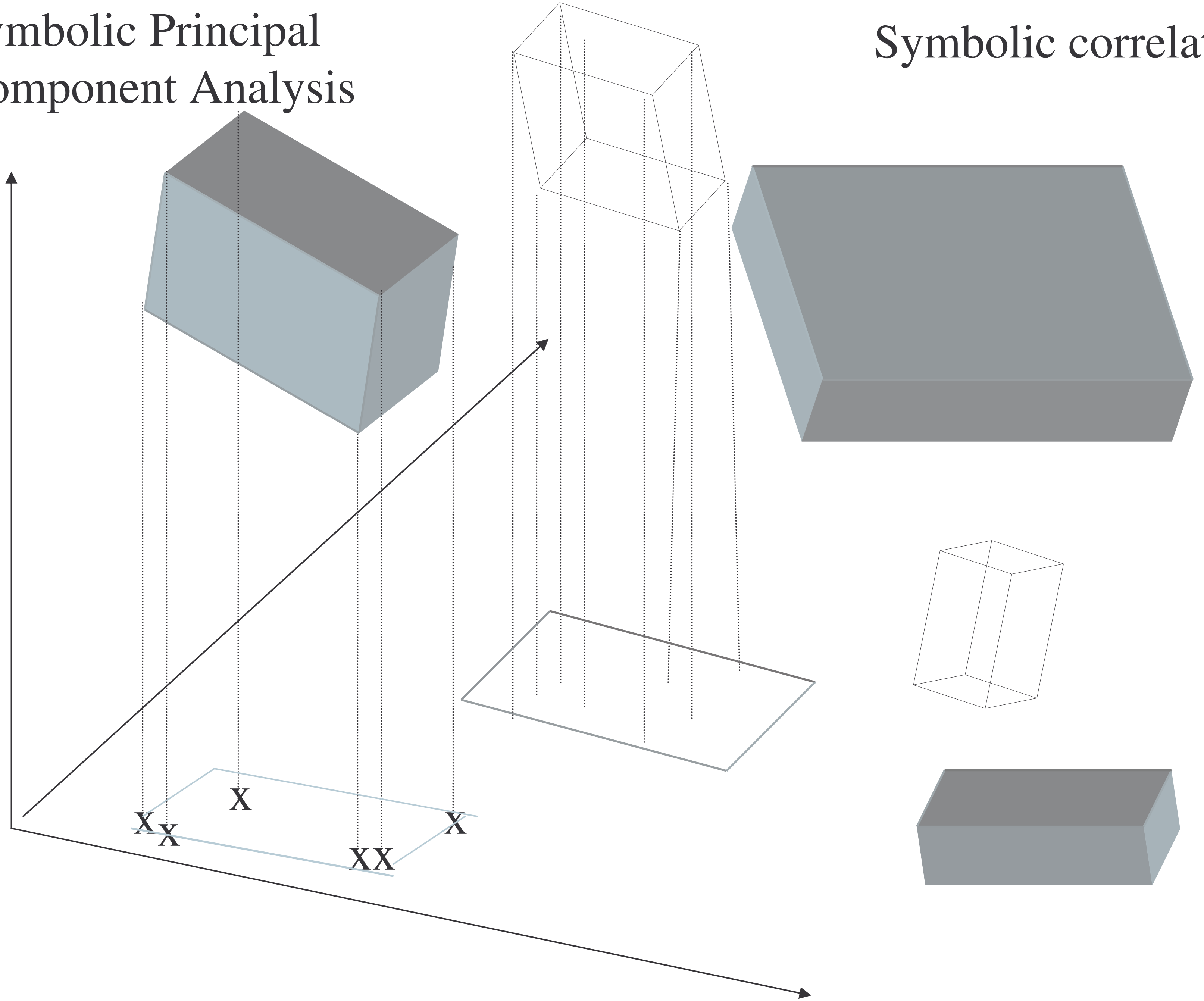


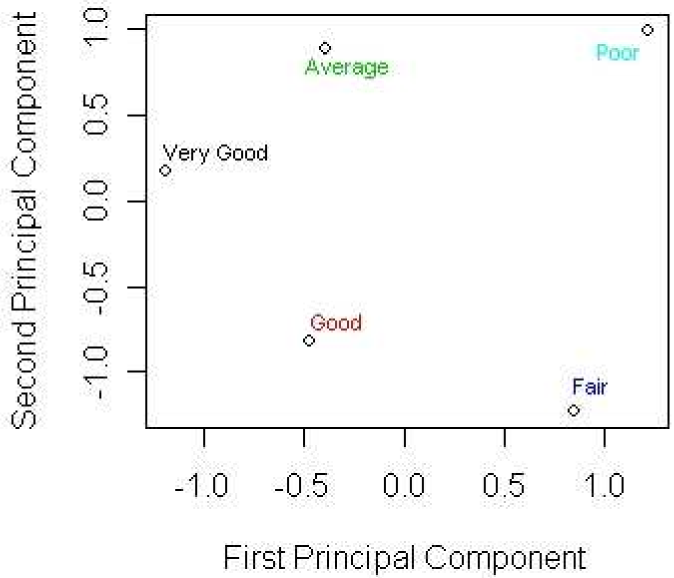
Variable à expliquer (le niveau de mortalité)

En **moins de branches**, **moins de nœuds** et avec une **meilleure discrimination**, l'arbre symbolique permet d'obtenir des classes de patients **plus homogènes** et **clairement expliquées** vis-à-vis de la variable « mortalité ».

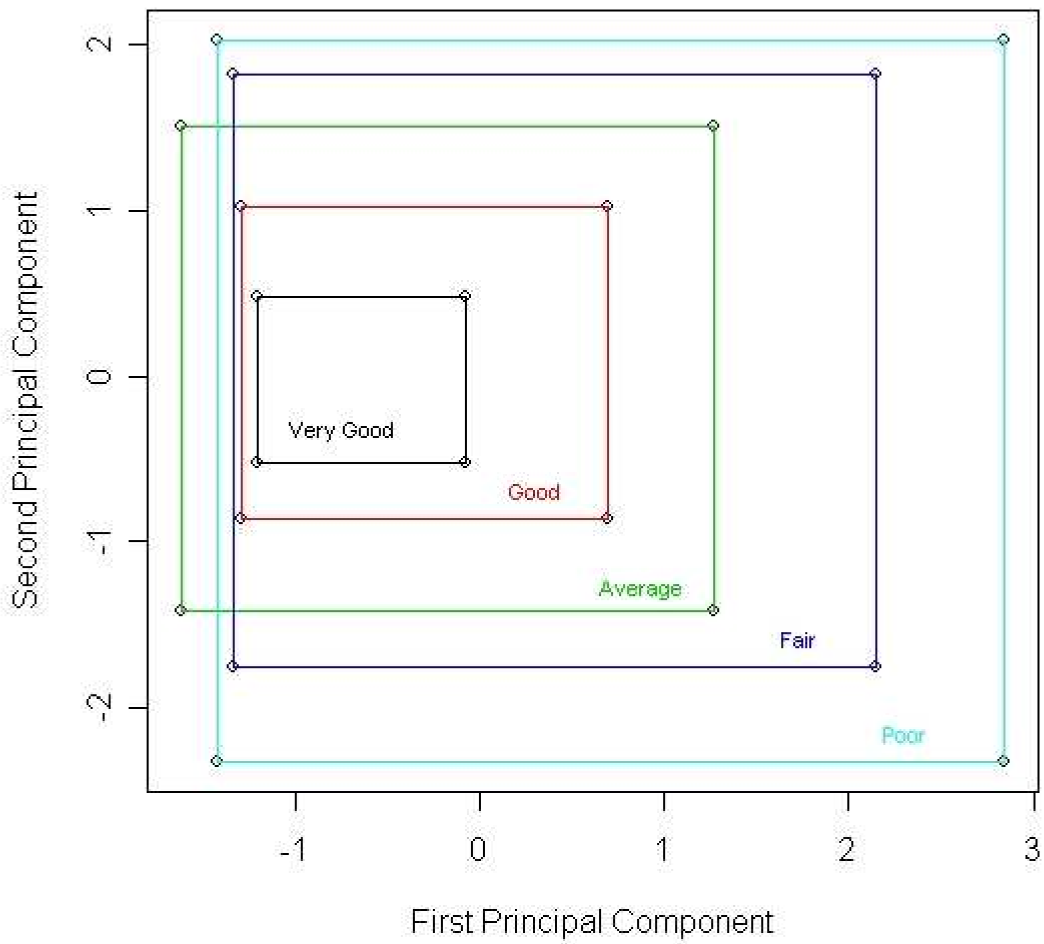
Symbolic Principal Component Analysis

Symbolic correlation





Classical Analysis



Symbolic Analysis

WHEN SYMBOLIC DATA ANALYSIS?

- When the good units are the concepts: finding why a team is a winner is not finding why a player is a winner
- When the categories of the class variable to explain are considered as new units and described by explanatory symbolic variables.
- When the initial data are composed by multisource data tables and then their fusion is needed

FRANCE IS DIVIDED INTO 50 000 COUNTIES CALLED IRIS

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

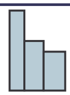
Classical Data table

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

Classical Data table

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

Symbolic description of London by the household variables

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

Symbolic description of London by the school variables

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

Concatenation

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

Adding Data to a SYMBOLIC FILE

Example: Social Security Insurance

315.000 Rows (actions)

Action	Provider	Medical act	Refund rate
c1000	B1	dentaire	65
c1001	B1	dental	100
c1002	B1	optique	65
c1003	B2	optical	35
c1004	B2	dentaire	65

A

Construction des données Symboliques, concept : « bénéficiaires»

Agrégation : Tableau à analyser conservant toute la richesse de l'information sous forme d'histogramme

19.000 Rows (Providers)

Provider	Medical act	Refund rate
B1	2/3 Dentaire, 1/3 optique	2/3 65, 1/3 100
B2	1/2 pharmacie, 1/2 dentaire	1/2 35, 1/2 65
....		

B

+

Provider	Gender	Age
B1	M	25
B2	F	64
...		

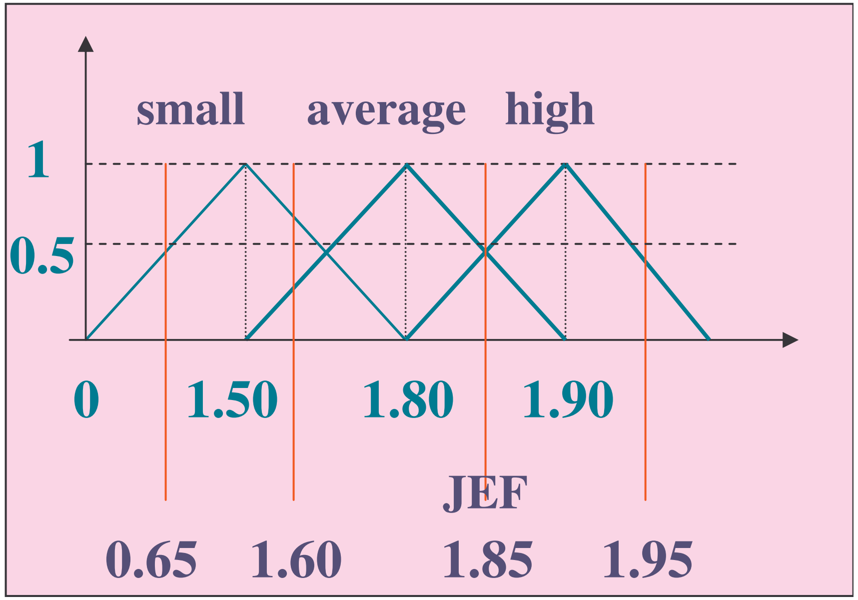
C

FROM FUZZY DATA TO SYMBOLIC DATA

	height	weight	hair
Paul	1.60	45	yellow
Jef	1.85	80	yellow
Jim	0.65	30	black
Bill	1.95	90	black

Initial Data

From Numerical to Fuzzy Data



	height			weight	hair
	small	average	high		
Paul	0.70	0.30	0	45	yellow
Jef	0	0.50	0.50	80	yellow
Jim	0.50	0	0	30	black
Bill	0	0	0.48	90	black

Fuzzy Data

	height			weight	hair
	small	average	high		
{Paul, Jef }	[0, 0.70]	[0.30, 0.50]	[0, 0.50]	[45, 80]	yellow
{Jim, Bill}	[0, 0.50]	0	[0, 0.48]	[30, 90]	black

Symbolic Data

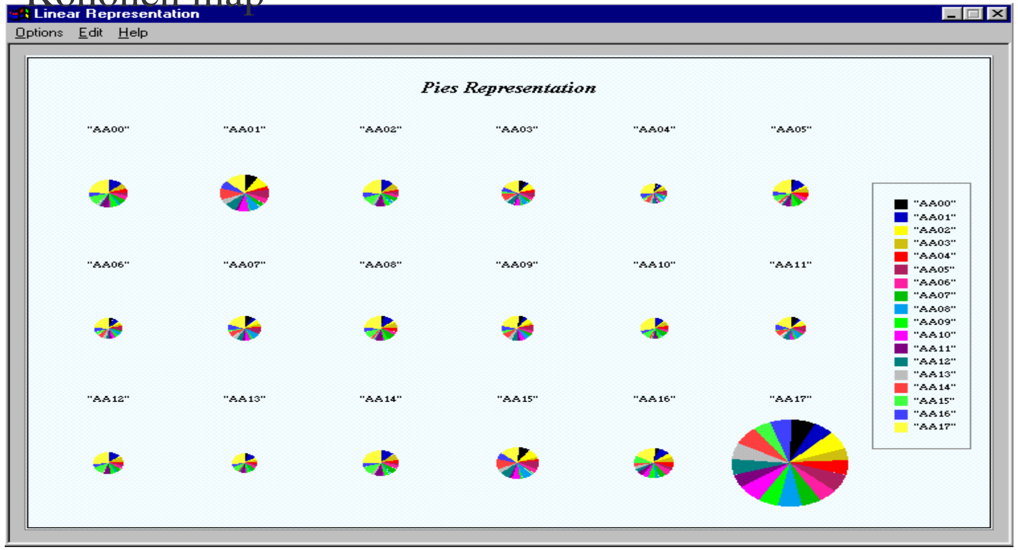
SOFTWARE COMPATIBLE WITH THE INPUT AND OUTPUT .SYR FILES:

SODAS SOFTWARE (2003)

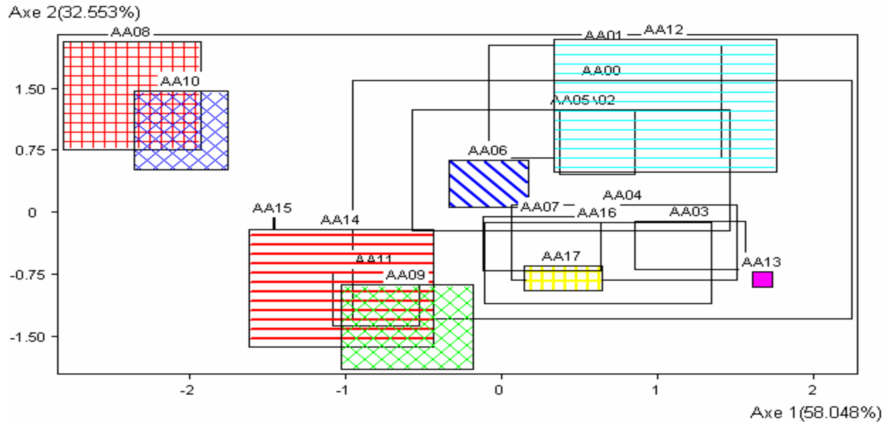
- SOE: symbolic objects edition.
- VIEW: Star graphics of symbolic objects
- DIV: Divisive clustering
- SCLUST: Symbolic clustering
- SPYR : Symbolic hierarchy and pyramid
- SOM: Kohonen mapping of interval variables
- SPCA: Principal Component Analysis for interval variables
- TREE: Symbolic decision tree.
- DISS: Dissimilarities between symbolic objects.

Examples of SDA Output

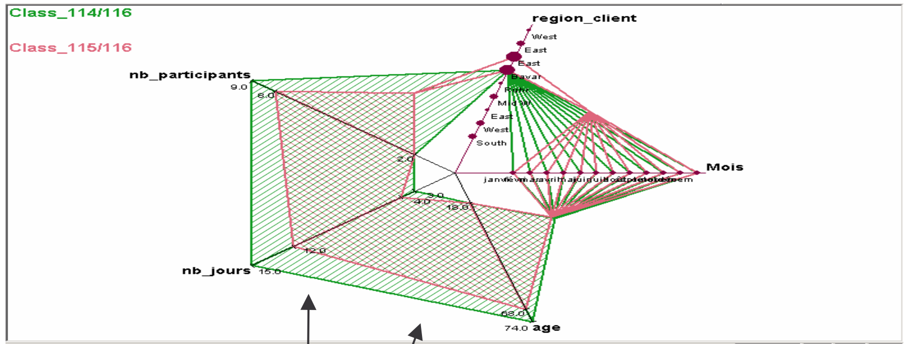
Kohonen map



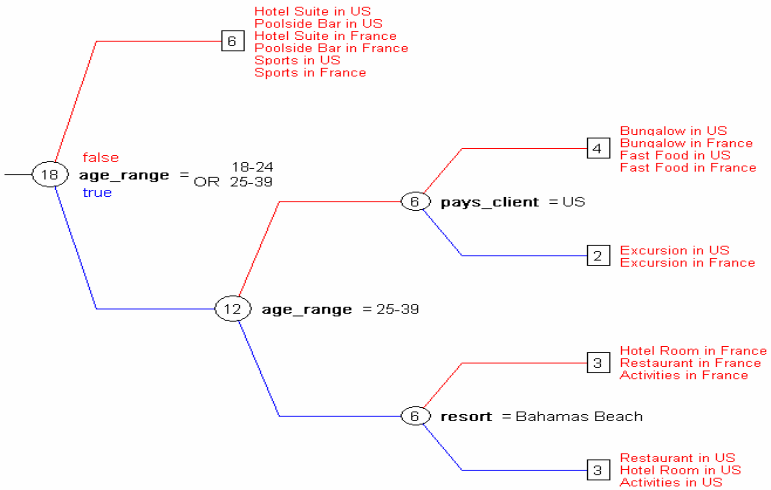
Principal componnent



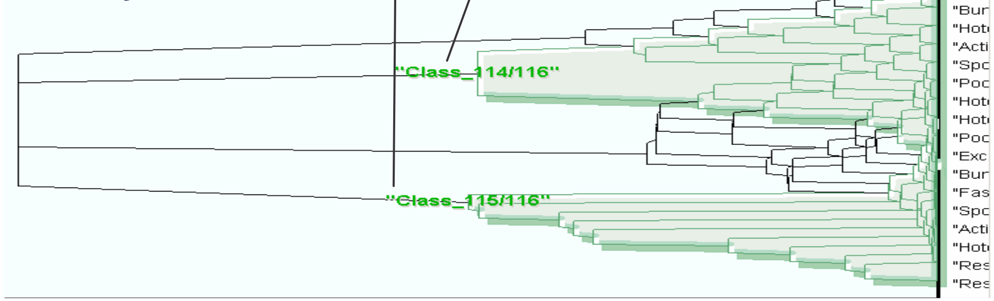
Zoom stars overlapping



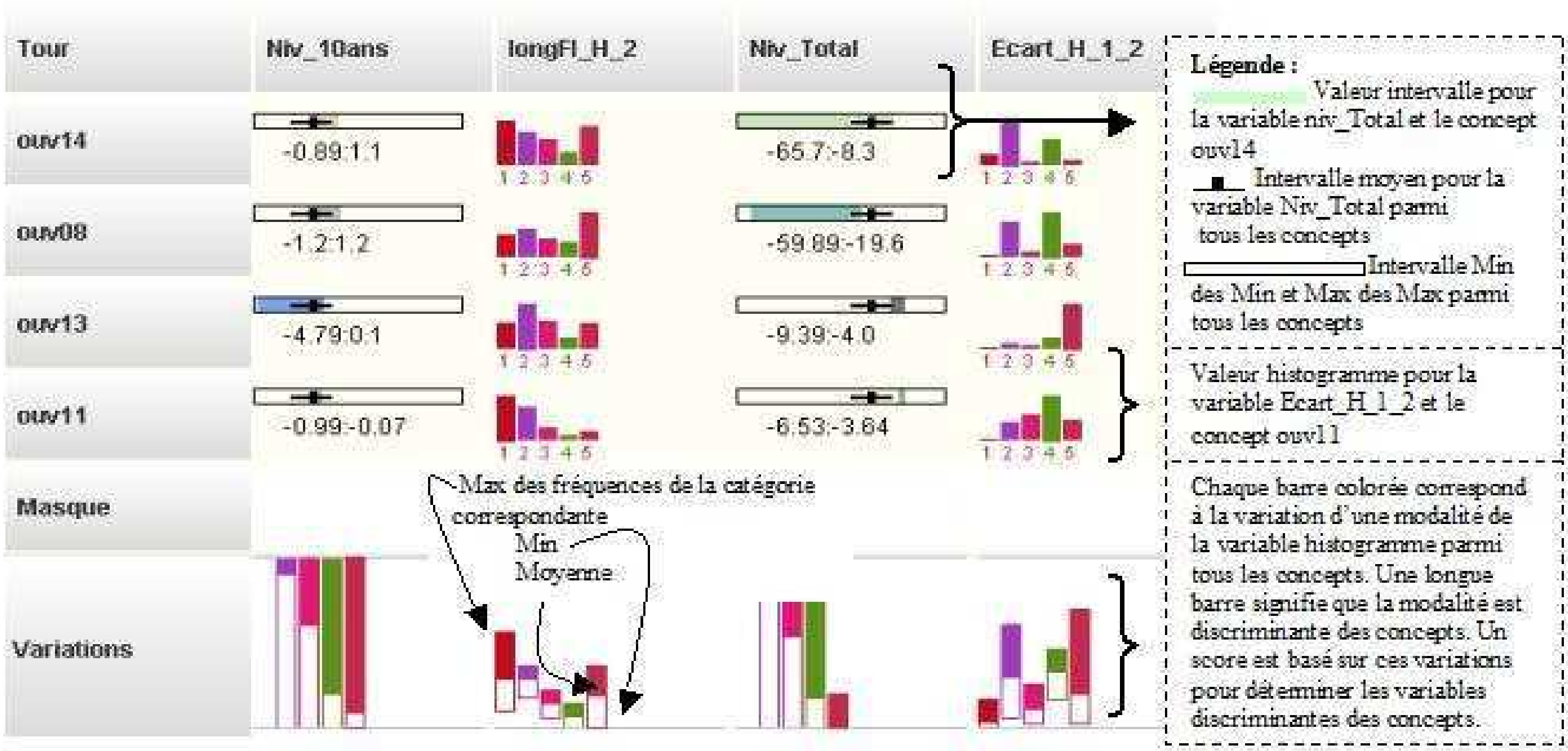
Top down clustering tree



Pyramid

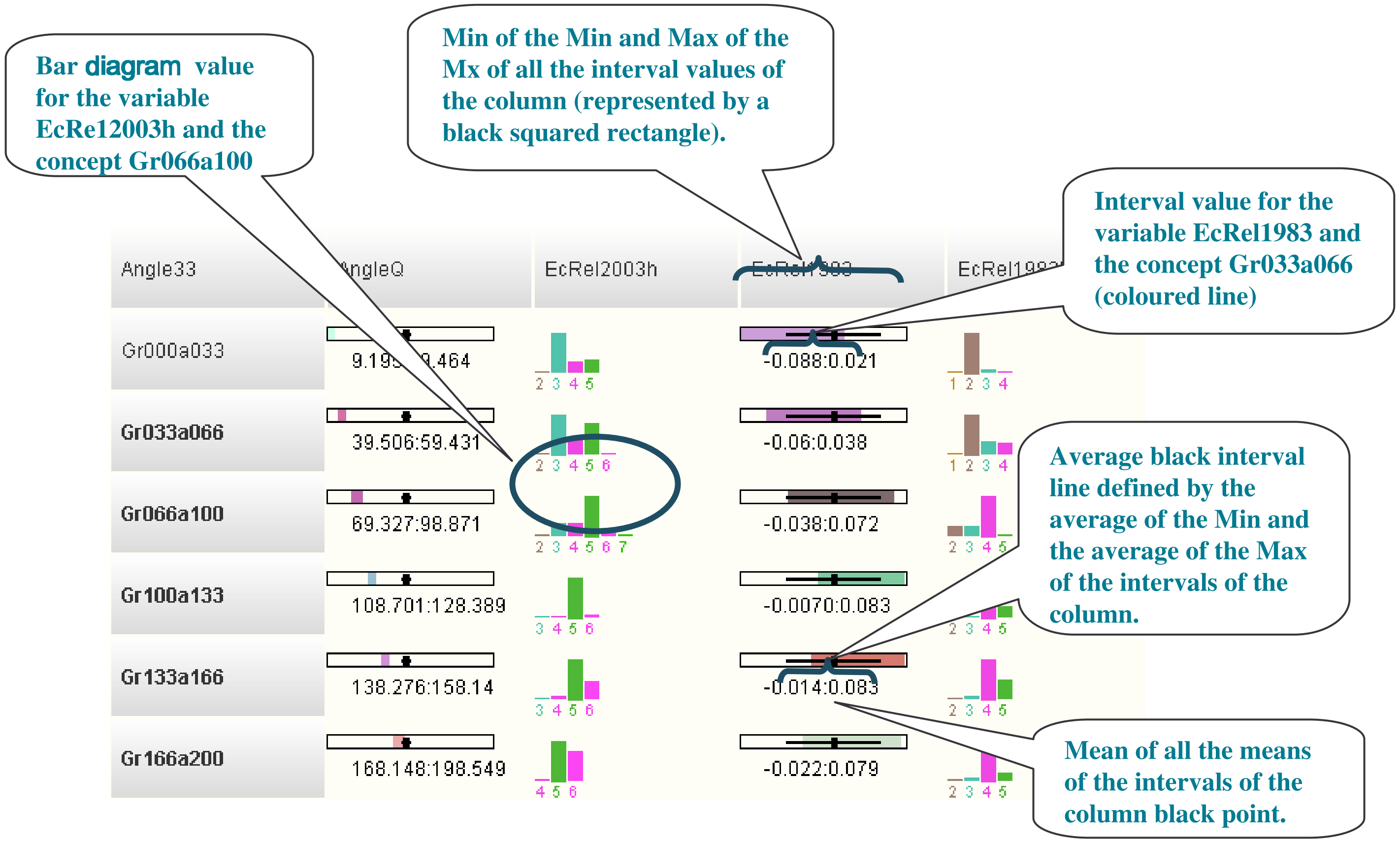


Management of Symbolic Data Table

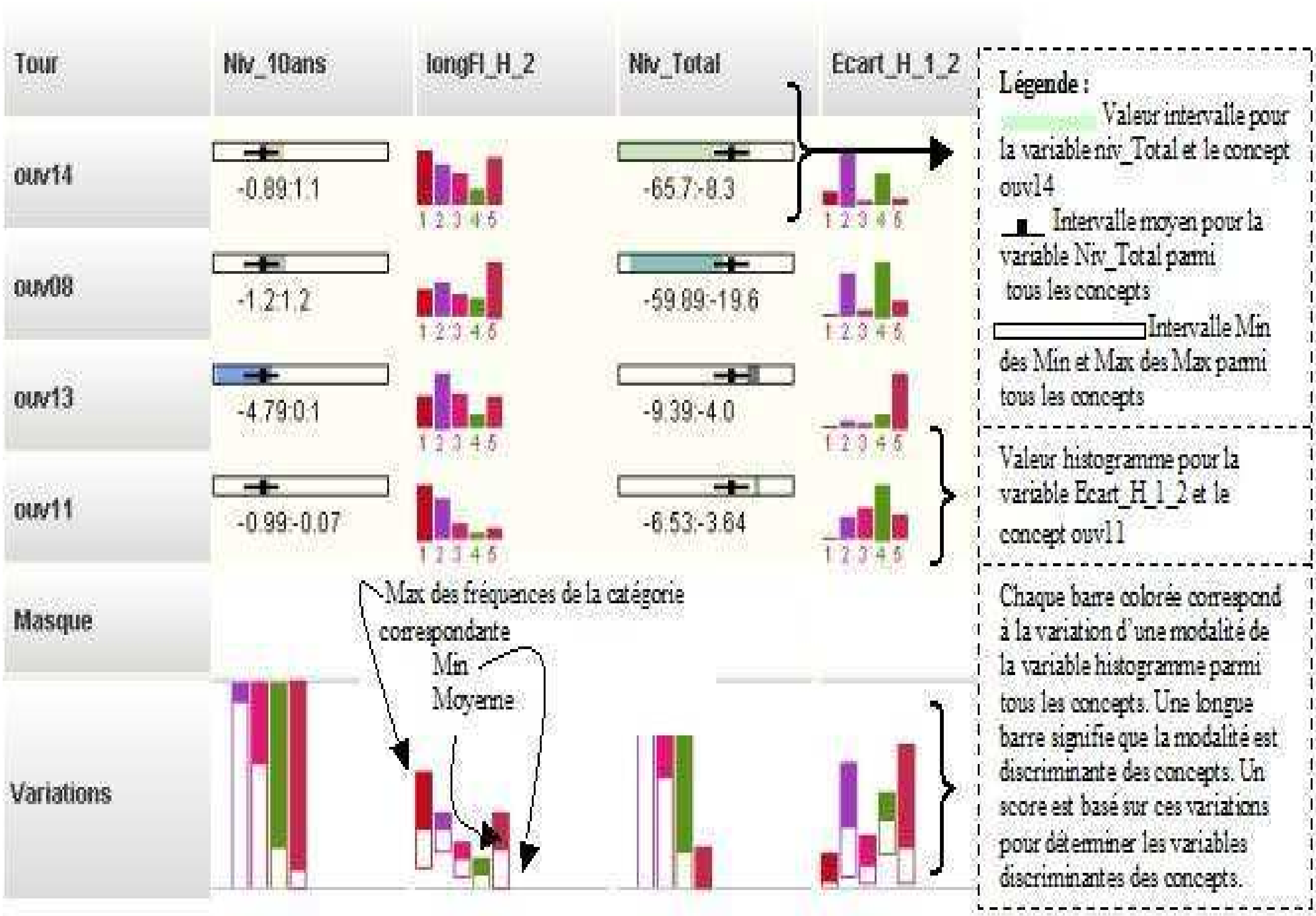


“Symbolic EXCEL”

Scoring the units is possible by min , max of the intervals or group of categories of the bar diagrams .



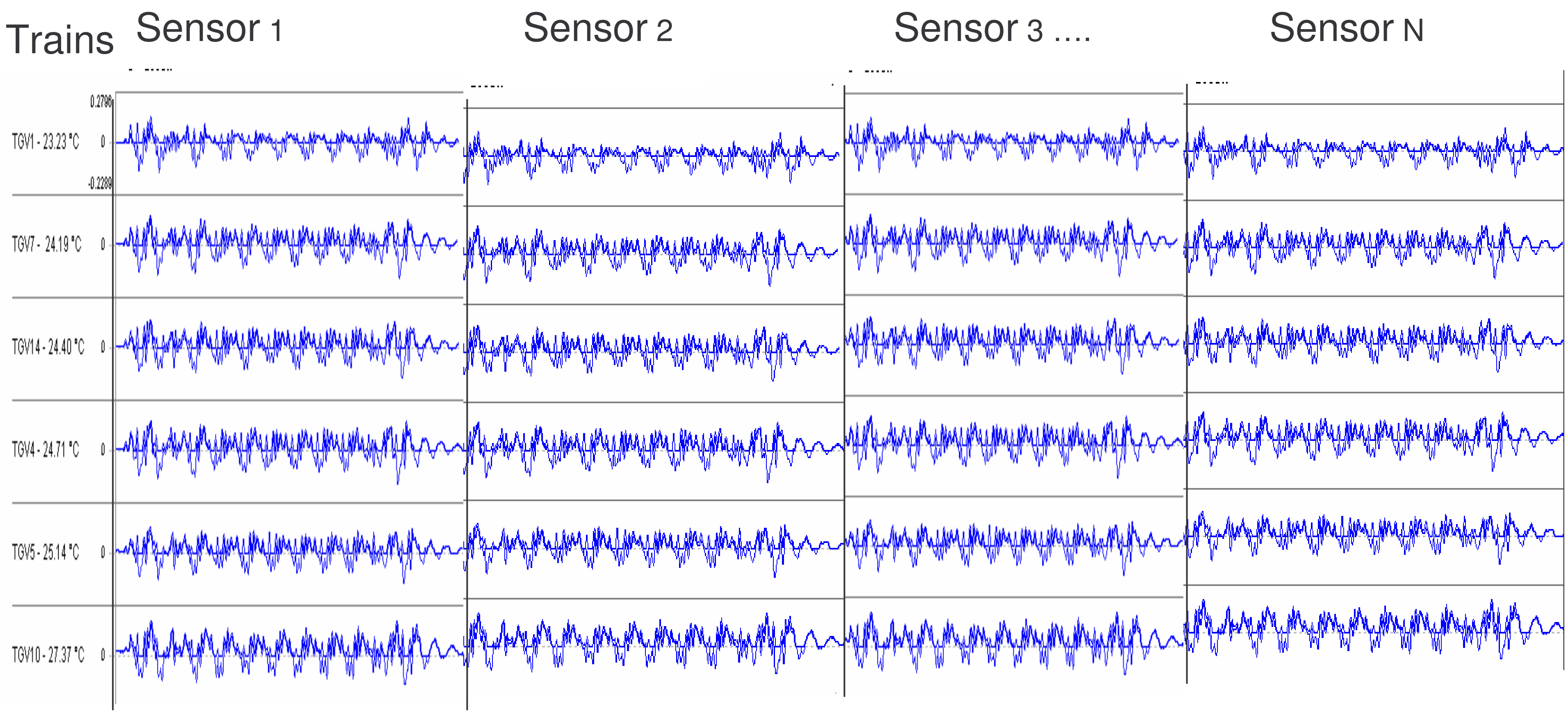
Scoring variables is also possible in order to select the most discriminate variables of the concepts :



Symbolic data analysis applications

- Trains
- Power Plants
- Social Security insurances
- Tackle security problems in regions
- Biology
- Catalogue Building

Anomaly detection on a bridge
(LCPC) Laboratoire Central Des Ponts et Chaussées



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

each cell contains until 800.000 values.

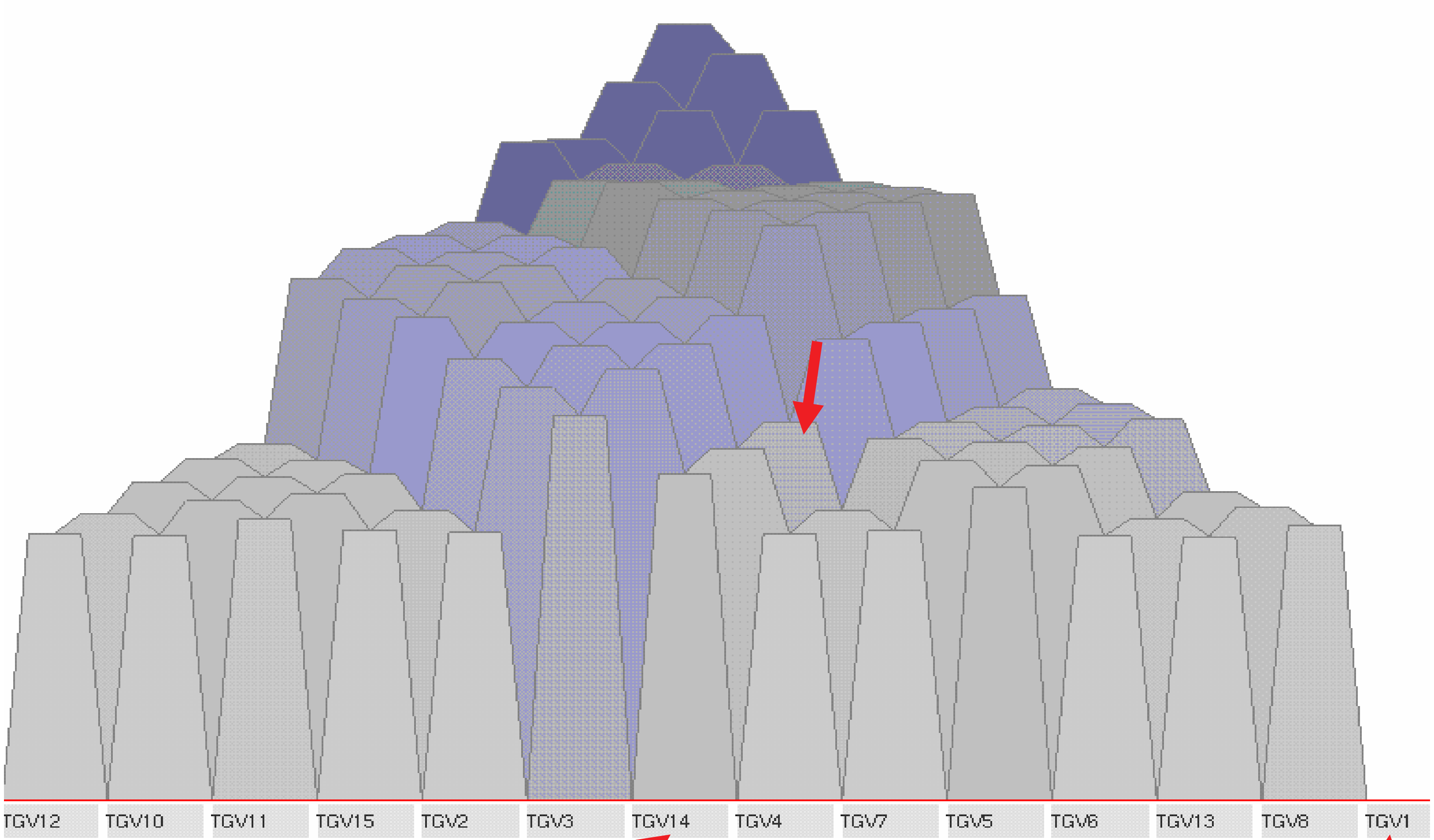
Each cell is transformed in HISTOGRAM from a PROJECTION or from WAVELETS

SYMBOLIC PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA on the interquartile intervals of the histograms contained in each cell

Two anomalies are easily detected: TGV1 is out of its group of temperature, TVG14 covers all the trains of its group of température

.



The symbolic pyramidal clustering confirm the anomalies.

1) TGV1 is out of its group of température

2) TGV 14 covers all the TGV of its group of température

NUCLEAR POWER PLANT

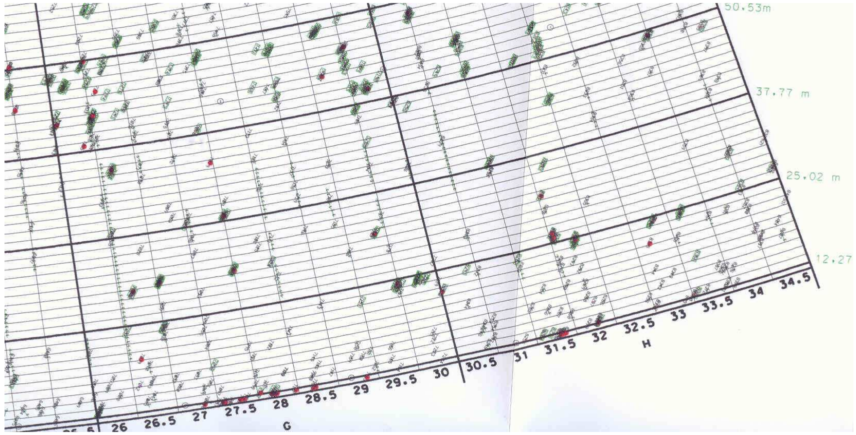
Nuclear thermal power station

Inspection :

Inspection machine

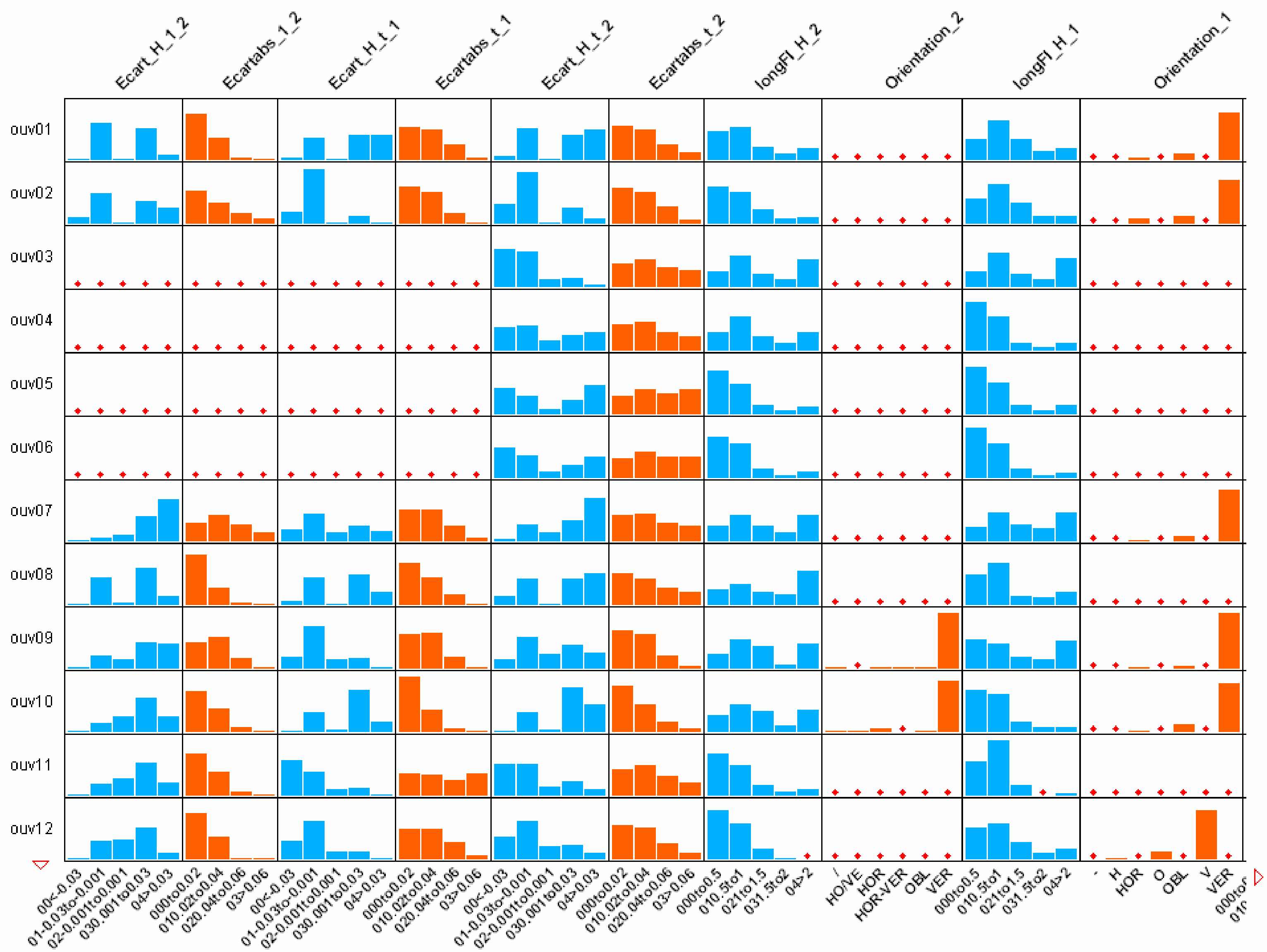
Craks

Cartography of the towel by a grid

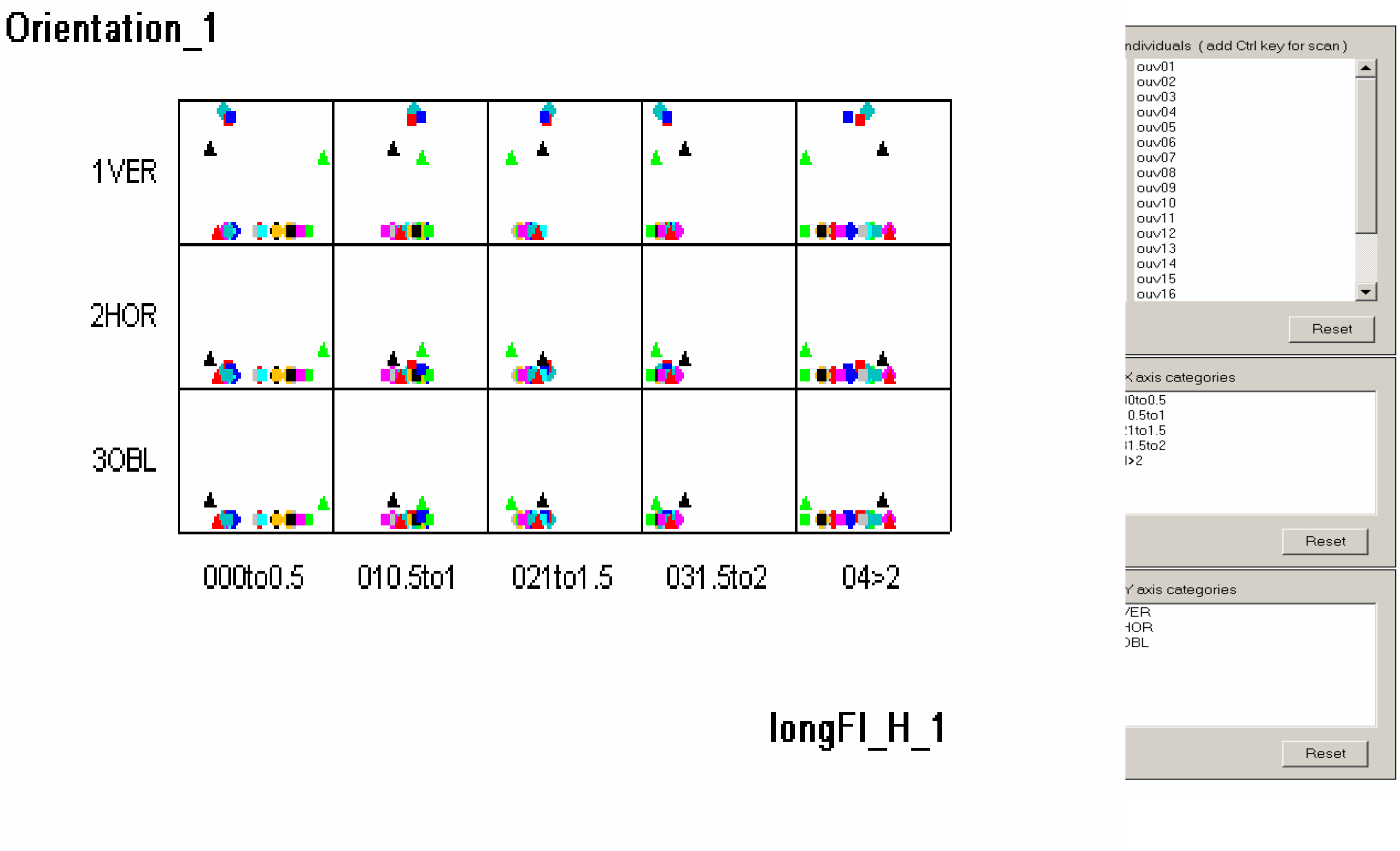


PB: FIND CORRELATIONS BETWEEN 3 CLASSICAL DATA TABLES OF DIFFERENT UNITS AND VARIABLES:
Table 1) Cracks description.
Table 2) Gap deviation of vertices of a grid at different periods compared to the initial model position.
Table 3) Gap depression from the ground.
ARE Transformed in ONE Symbolic Data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

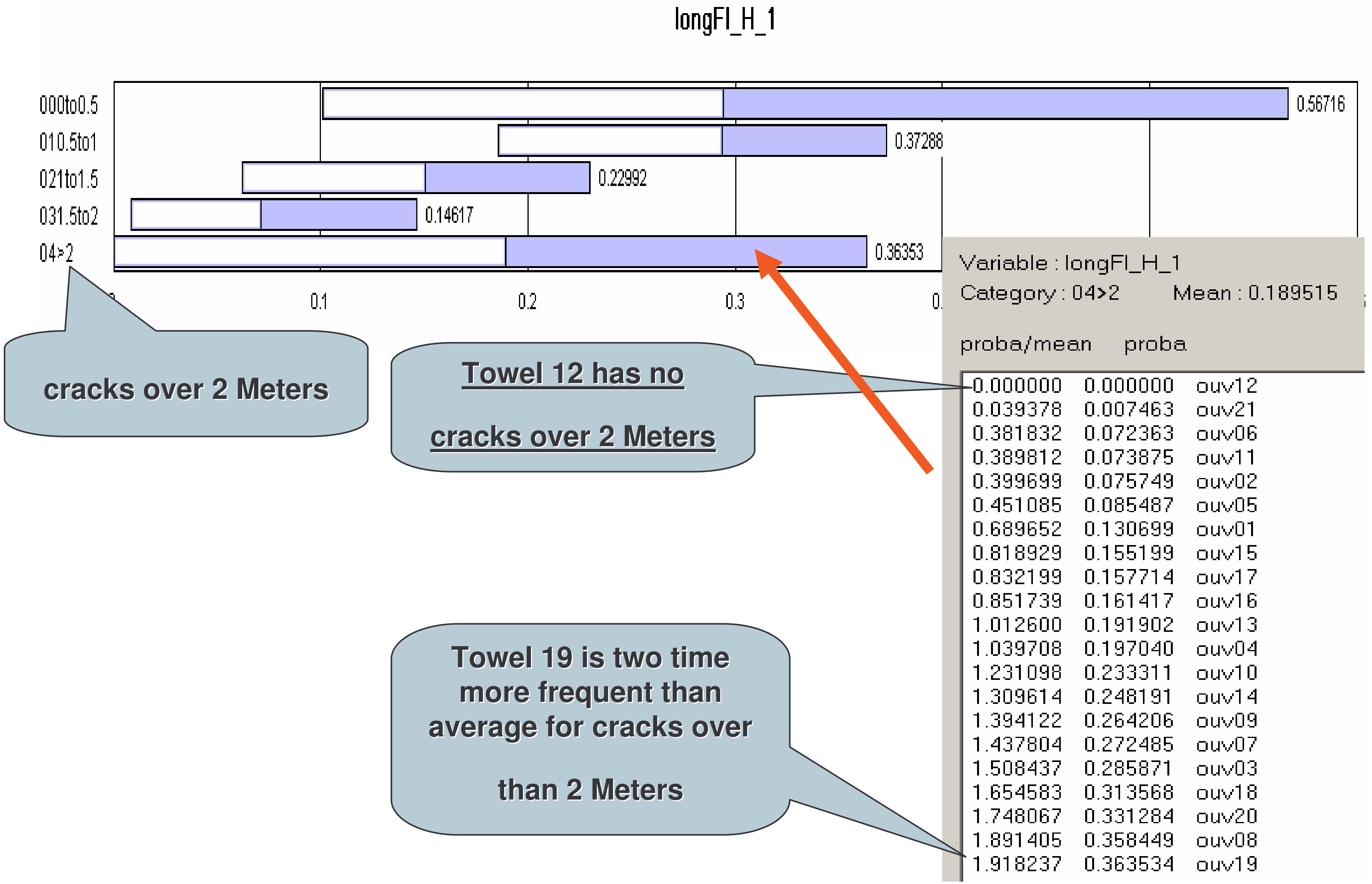
Symbolic Data Table from STATSYR



Crossing histograms by STATSYR

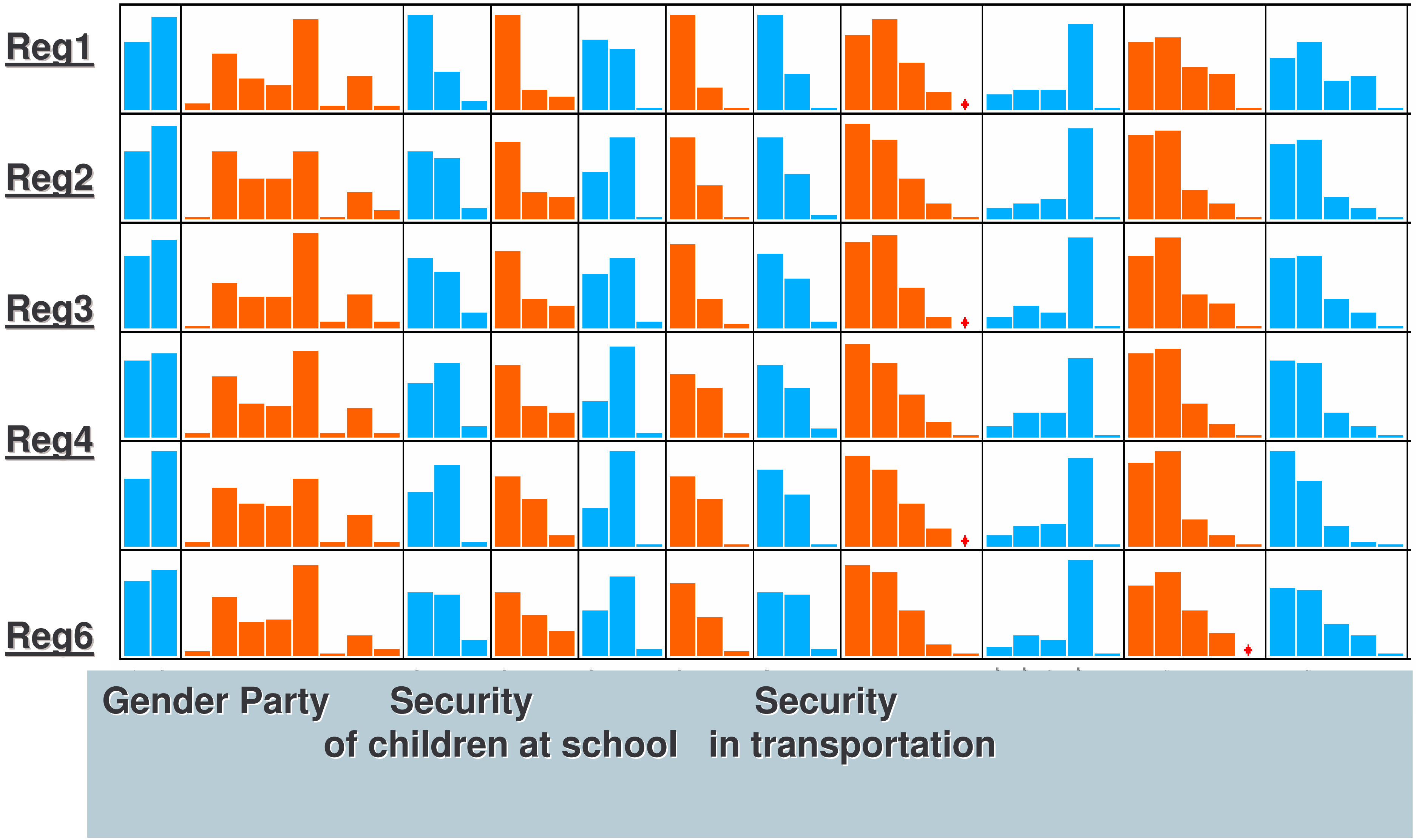


Cracks description

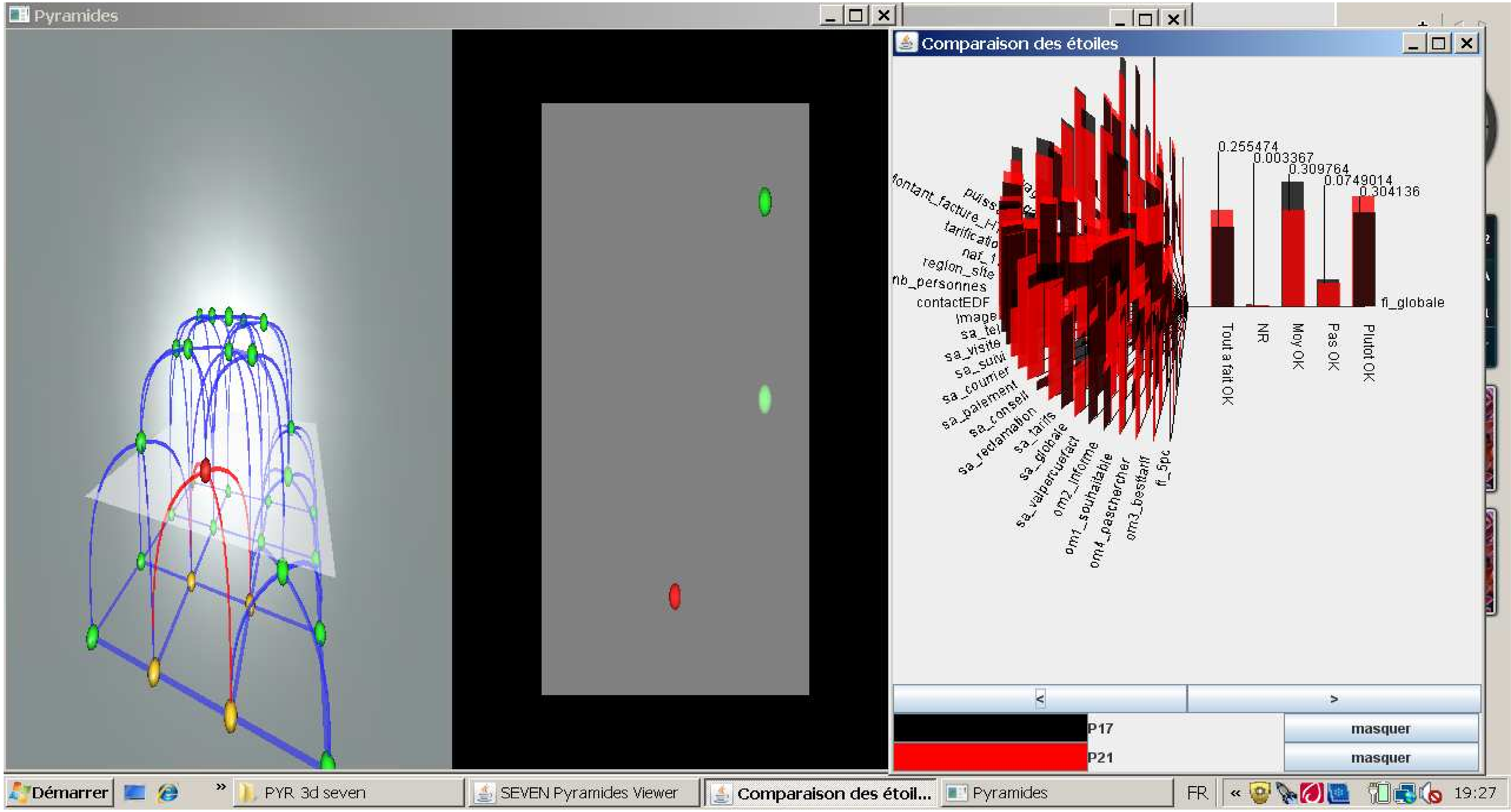


proba/mean	proba	
0.000000	0.000000	ouv12
0.039378	0.007463	ouv21
0.381832	0.072363	ouv06
0.389812	0.073875	ouv11
0.399699	0.075749	ouv02
0.451085	0.085487	ouv05
0.689652	0.130699	ouv01
0.818929	0.155199	ouv15
0.832199	0.157714	ouv17
0.851739	0.161417	ouv16
1.012600	0.191902	ouv13
1.039708	0.197040	ouv04
1.231098	0.233311	ouv10
1.309614	0.248191	ouv14
1.394122	0.264206	ouv09
1.437804	0.272485	ouv07
1.508437	0.285871	ouv03
1.654583	0.313568	ouv18
1.748067	0.331284	ouv20
1.891405	0.358449	ouv08
1.918237	0.363534	ouv19

Tackle security problems in regions



Symbolic Spatial Classification



Réalisé dans le cadre de l'ANR SEVEN (EDF, LIMSI, Dauphine).

Théorie de la classification spatiale: E. Diday (2008) “Spatial classification”. DAM (Discrete Applied Mathematics) Volume 156, Issue 8, Pages 1271-1294.

UNDERLYING MATHEMATICAL THEORIE

- THE SYMBOLIC VARIABLE VALUES ARE RANDOM VARIABLE .
- STOCHASTIC GALOIS LATTICES ARE THE ALGEBRAIC STRUCTURE OF SYMBOLIC OBJECTS (presented by G. Choquet, Acad of Sciences)
- THE COPULAS THEORY IS THE UNDERLYING PROBABILISTIC STRUCTURE OF SYMBOLIC OBJECTS

Future development

- **Mathematics:** it can be shown that the underlying structure of symbolic descriptions of concept are "stochastic Galois Lattices". New algebra is needed.
- **Statistics:** the underlying model of symbolic variables are variables whose values are random variables instead of numbers as usual. "Copulas" are needed. Much work is needed for validation, stability, robustness of the results.
- **Computer sciences:** extending data base to symbolic data bases , queries and language of the primitives. Extending EXCEL to SYMBOLIC EXCEL is done in the SYR software, much remains to be done.
- **Applications:** all domains where new knowledge has to be extracted from small or large data bases.

TWO SYMBOLIC DATA ANALYSIS SOFTWARES

- **SODAS (2003)**

FREE from 2 European Consortium

- **click : SODAS CEREMADE**

- **SYR (2008)**

More professional from SYROKKO Company

- **Click: www.syrokko.com**

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.

CONCLUSION

- If you have standard units described by numerical and (or) categorical variables, these variables induce categories which can be considered as new units called “concepts” described by symbolic variables taking care of their internal variation. Then SDA can be applied on these new units in order to get complementary and enhancing results by extending standard analyis to symbolic analysis.

SPATIAL CLASSIFICATION

Here the goal of a spatial classification is to position the units on a spatial network and to give simultaneously a set of homogeneous structured classes of these units "compatible with the network".

TAKE CARE ! SPATIAL CLASSIFICATION

IS NOT CLASSICATION OF SPATIAL DATA.

SPATIAL PYRAMIDAL CLUSTERING

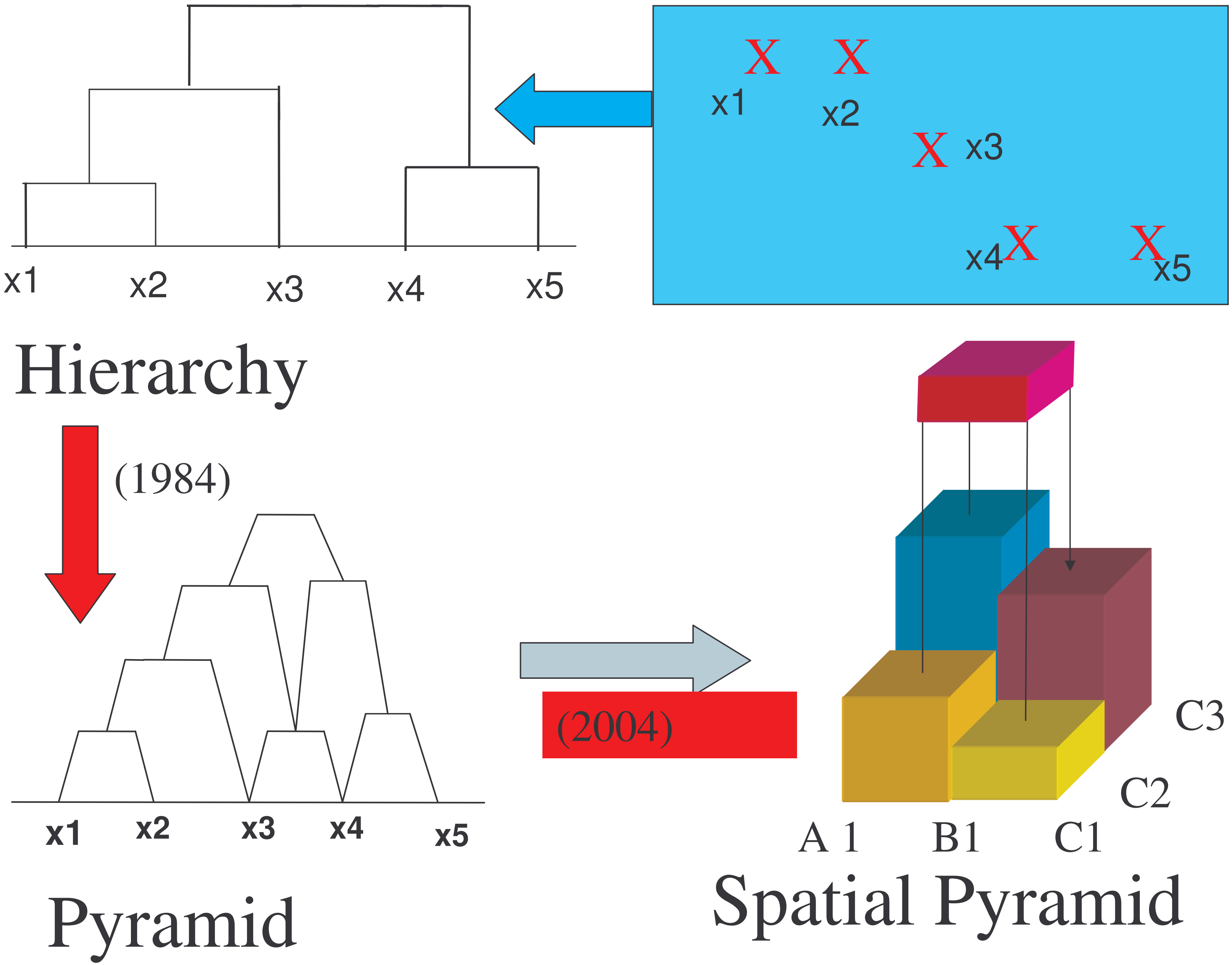
Instead of representing the clusters associated to each level of a standard hierarchical or pyramidal clustering on an ordered line our aim is to represent them on a surface or on a volume .

GOAL

Extending standards hierarchies and pyramids

TO

Spatial hierarchies and spatial pyramids such that each cluster be a convex of a spatial network


x1
x2
x3
x4
x5
X x1
X x2
X x3
x4 X
X x5
Hierarchy
(1984)
x1
x2
x3
x4
x5
Pyramid
(2004)
C3
C2
A 1
B1
C1
Spatial Pyramid

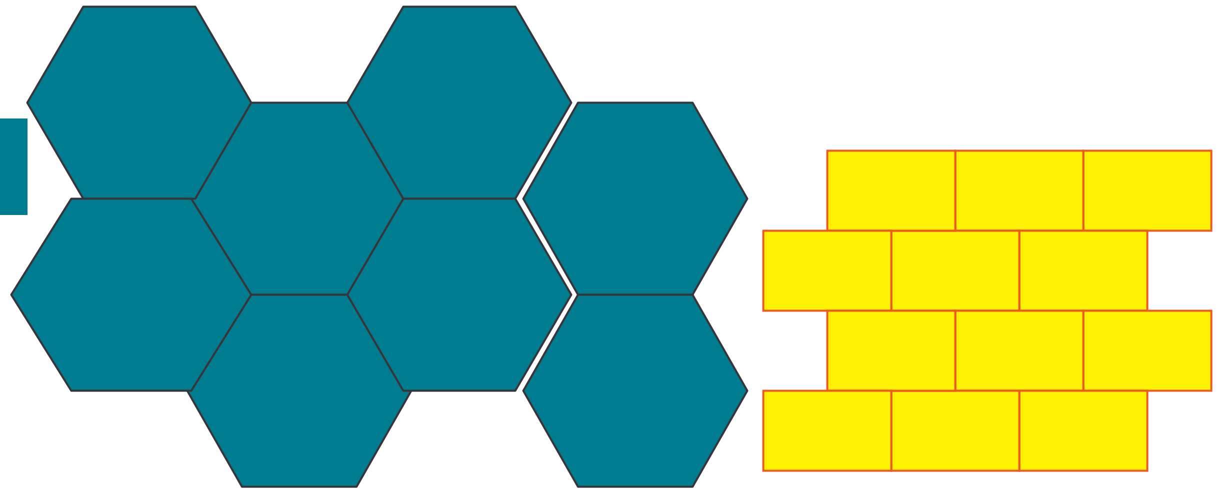
WHAT IS A (m, k)- network ?

IT IS A GRAPH WHERE:

i) m arcs defining m equal angles, meet at each node.

ii) smallest cycles contain k arcs of equal length.

(3,6)-network



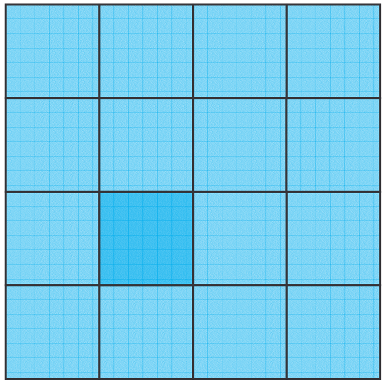
A (m, k)-network is a tesselation but a tesselation is not necessarily an (m,k) network

There are only three (m-k)-networks

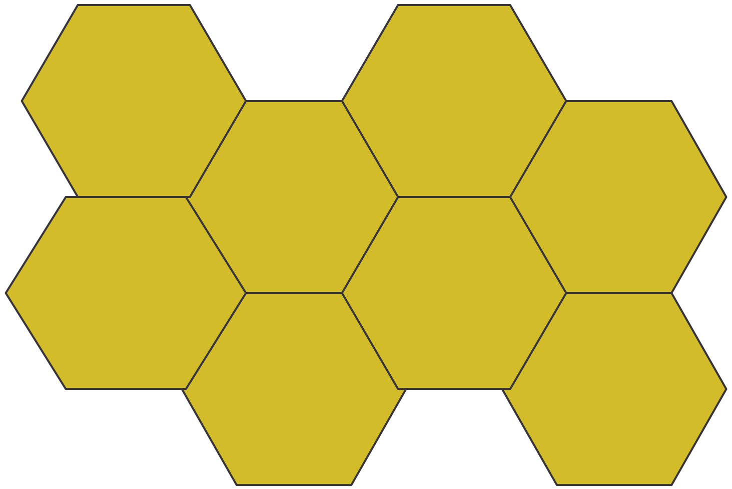
(m,k) = (3,6) where the cells are hexagones,

(m,k) = (4,4) where the cells are square : a grid

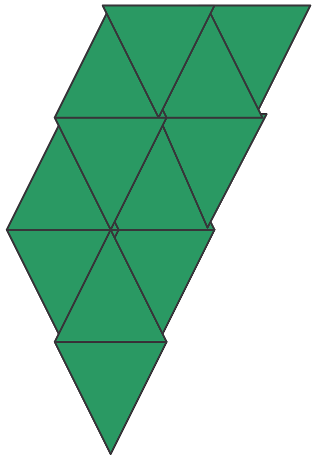
(m,k) = (6,3) where the cells are equilateral triangles .



(4,4)-network

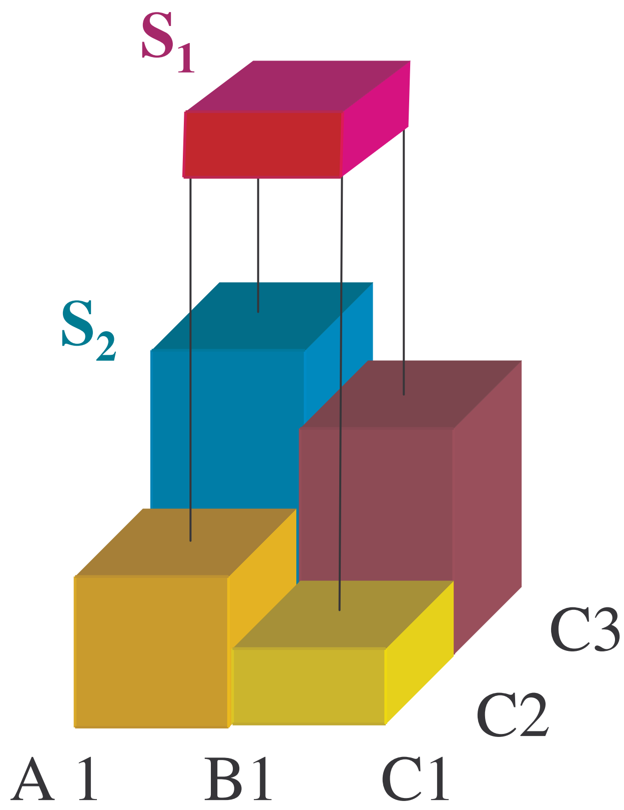


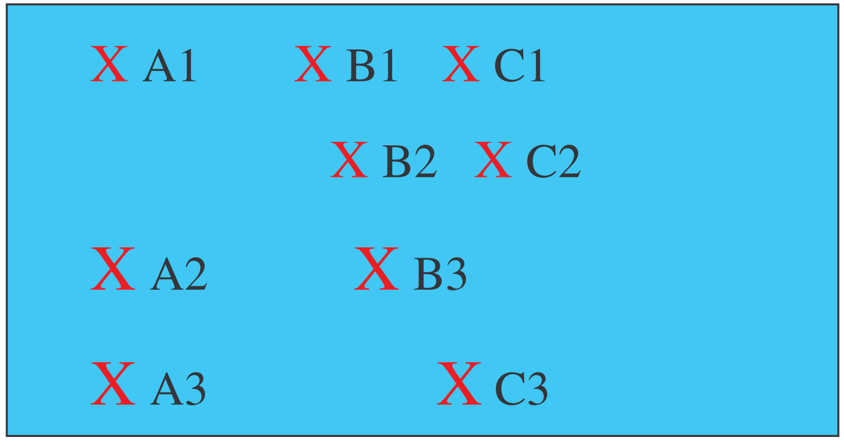
(3,6)-network

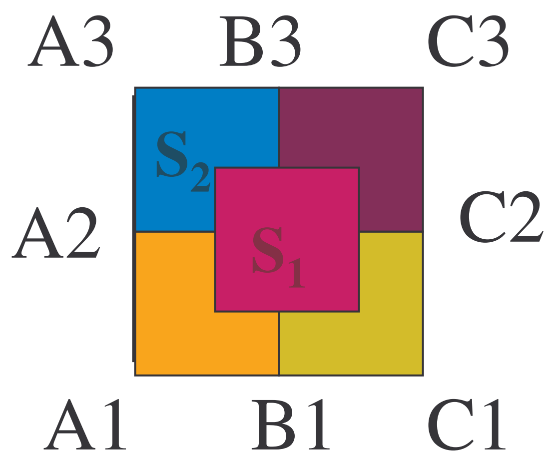


(6,3)-network

EXAMPLE OF SPATIAL PYRAMID

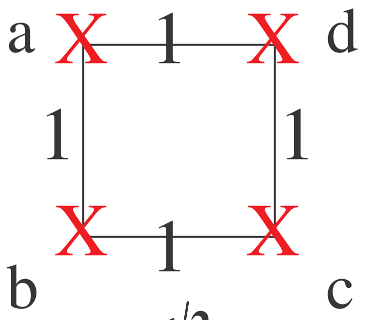






SPATIAL PYRAMID OF 9 UNITS ON A (4,4)-NETWORK
CLASSES OVERLAP: B1 BELONGS IN 2 CLASSES.

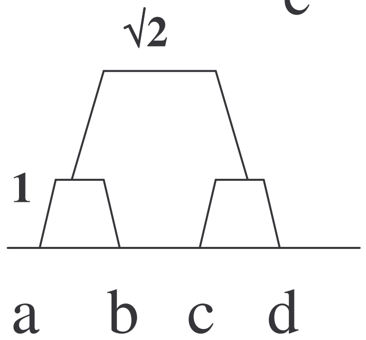
Initial Data



	a	b	d	c
a	0	1	√2	√2
b		0	2	√2
d			0	1
c				0

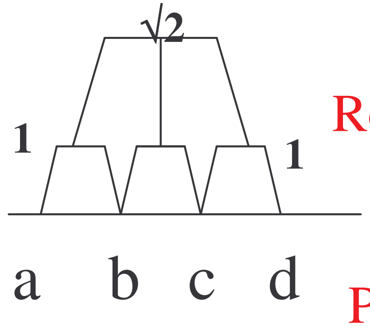
Ultrametric

Hierarchy



Robinsonienne

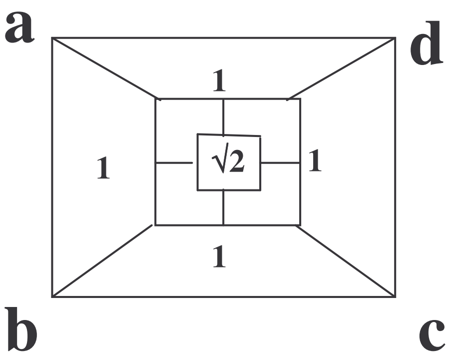
Pyramid



	a	b	d	c
a	0	1	√2	√2
b		0	2	1
d			0	1
c				0

Spatial pyramid

Yadidean



	a	b	d	c
a	0	1	1	√2
b		0	√2	1
d			0	1
c				0

With only 2 levels we get a better fit with the initial distance!!!

Definition of a "d-grid matrix"

$W(d) = \{d(x_{ik}, x_{jm})\} \quad i, j \in \{1,..,p\}, k, m \in \{1,..,n\}^{\bullet}$

Where x_{ij} is a vertice of the grid.

Definition of a Robinsonian Matrix

We recall that a Robinsonian matrix is symmetrical, its terms increase in row and column from the main diagonal and the terms of this diagonal are equal to 0.

Definition of a "Robinsonian by blocks matrix“

It is a d-grid block matrix $Z(d)$ such that:

i) it is symmetrical,

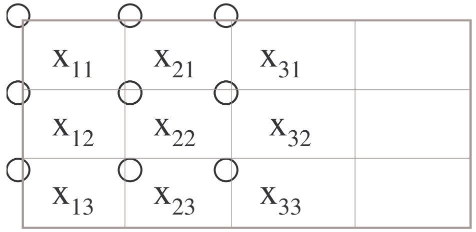
ii) the matrices of its main diagonal $Z_{ii}(d) = X_i X_i^T(d)$ are Robinsonian.

iii) The matrices $Z_{ij}(d) = X_i X_j^T(d)$ are symmetrical and increase in row and column from the main diagonal.

Definition of a “Yadidean matrix“

A d-grid matrix $Y(d) = \{d(x_{ik}, x_{jm})\}_{i,\, j\in \{1,\dots p\},\, k,\, m\in \{1,\dots n\}}$, induced by a grid M is Yadidean, when the d-grid blocks matrix $Z(d) = \{X_i X_j^T(d)\}_{i,\, j\in \{1,..,p\}}$ induced by M is Robinsonian by blocks.

The d_M dissimilarity induced from the grid.

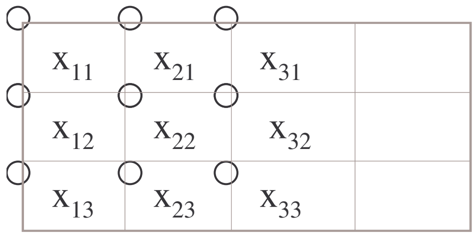


A 3x3 Grid

	x_{11}	x_{12}	x_{13}	x_{21}	x_{22}	x_{23}	x_{31}	x_{32}	x_{33}
x_{11}	0	1	2	1	2	3	2	3	4
x_{12}		0	1	2	1	2	3	2	4
x_{13}			0	1	2	1	4	4	2
x_{21}				0	1	2	1	2	3
x_{22}					0	1	2	1	2
x_{23}						0	3	2	1
x_{31}							0	1	2
x_{32}								0	1
x_{33}									0

IT IS A ROBINSON BY BLOCKS MATRIX

A YADIDEAN DISSIMILARITY



A 3x3 Grid

$$X_2\, X_3^T(d) = \begin{vmatrix} 1 & 1 & 8 \\ 1 & 1 & 3 \\ 8 & 3 & 1 \end{vmatrix}$$

	x_{11}	x_{12}	x_{13}	x_{21}	x_{22}	x_{23}	x_{31}	x_{32}	x_{33}
x_{11}	0	4	8	4	4	7	5	8	8
x_{12}		0	5	4	4	5	8	7	6
x_{13}			0	7	5	5	8	6	6
x_{21}				0	1	4	1	1	8
x_{22}					0	3	1	1	3
x_{23}						0	5	3	1
x_{31}							0	1	8
x_{32}								0	3
x_{33}									0

The upper part of a Yadidean matrix Y(d) of a 3x3 grid and the block matrix $X_2\, X_3^T(d)$ of its associated Robinsonian by blocks matrix.

PROPERTIES OF A YADIDEAN MATRIX

A Yadidean matrix is not Robinsonian, as its terms : the $d(x_{ik}, x_{jm})$ for $i, j \in \{1,\ldots,p\}$ and $k, m \in \{1,\ldots,n\}$) do not increase in row and column from the main diagonal

The maximal percentage of different values in a Yadidean matrix among all possible dissimilarities is

$$x = K(n, p)\ 200/\ np\ (np-1) = 50 + 100\ (n+p-2)/\ 2(np-1)$$
$$x = 100\ K(n, n)\ (\ 2/\ n^2\ \ (n^2-1)) = 50 + 100/(n+1) \text{ when } p = n.$$

THEREFORE THE MAXIMAL PERCENTAGE OF DIFFERENT VALUES TENDS TO BE TWO TIME LESS THEN IN A DISSIMILARITY OR A ROBINSON MATRIX.

THE NUMBER OF CLASSES IN A CONVEX PYRAMID TENDS TO BECOME TWO TIMES LESS THAN IN A STANDARD PYRAMID

COMPATIBILITY BETWEEN A DISSIMILARITY AND A GRID

A dissimilarity d is "diameter conservative" for M when for any convex C of M we have

$D(C, d_M) = d_M(i, k) \Rightarrow D(C, d) = d(i, k)$.

In this case we say that d is "compatible" with M.

Proposition

A dissimilarity is compatible with a grid if and only if it is Yadidean.

OVERVIEW ON ONE TO ONE CORRESPONDENCES

Hierarchies	$\Longleftrightarrow$	Ultrametrics
$\cap$		$\cap$
Pyramids	$\Longleftrightarrow$	Robinsonian
$\cap$		$\cap$
Spatial Convex Pyramids	$\Longleftrightarrow$	Yadidean

WHY THESE BIJECTIONS ARE IMPORTANT ?

D = THE GIVEN INITIAL DISSIMILARITY

D’ = Yadidean-dissimilarity ⟷ A SPATIAL PYRAMID

THE DISTORSION BETWEEN D and the S-PYRAMID

IS

THE DISTORSION BETWEEN D and D’.

Definition of a spatial pyramid

A spatial pyramid on a finite set Ω is a set of subsets (called “class") of Ω satisfying the following conditions :

1) $\Omega \in P$
2) $\forall\, w \in \Omega$, $\{w\} \in P$.
3) $\forall\, (h, h') \in P \times P$ we have $h \cap h' \in P \cup \varnothing$
4) *There exists a m/k-network of Ω such that each element of P is convex, connected or maximal.*

Definition of a standard pyramid

4) *There exists an order for which each class is an interval .*

Building a Spatial Pyramid

- **1) .Each element of Ω is considered as a class and added to P.**
- **2). Each mutual neighbor classes which can be merged in a new convex, among the set of classes already obtained and which have not been merged four times, are merged in a new class and added to P.**
- **3). The process continues until all the elements of Ω have been merged.**

During the process:

- - Each time a new convex is created an order is fixed for its rows and columns.
- - Two convexes cannot be merged if they are not connected.
- - A convex C' which is contained in another convex C and which does not contain a row or a column of the border of C, cannot be aggregated with any convex external to C.
- This algorithm can be applied to any kind of dissimilarity and aggregation index.
- By deleting all the classes which are not intersections of two different classes of P the algorithm SCAP produces a weakly large spatial pyramid (P, f).

Different kinds of convexes induced by a Yadidean dissimilarity

Definition of a "maximal (M, d)-convex“

- A convex C of M is called a "maximal (M, d)-convex" if there is not a convex C' of M such that $C \subset C'$ (strictly) and D(C', d) = D(C, d).
- In a Yadidean matrix Y = {d(xik, xjm)}i,j∈{1,…p}, k,m∈{1,…n}, such a convex C is easy to find as it is characterized by the fact that if its diameter is D(C, d) = d(xik, xjm) and if i<j and k<m, then, the same value does not exist:
- in any row or column smaller than k and higher than m if i and j are fixed (i.e. among the terms d(xik', xjm') where k'≤k and m'≥m in the matrix XiXjT(d)),
- in any row or column lower than i and higher than j if k and m are fixed (i.e. among the matrices Xi'Xj'T(d) with i'≤ i and j'≥j).

Indexed Spatial pyramid

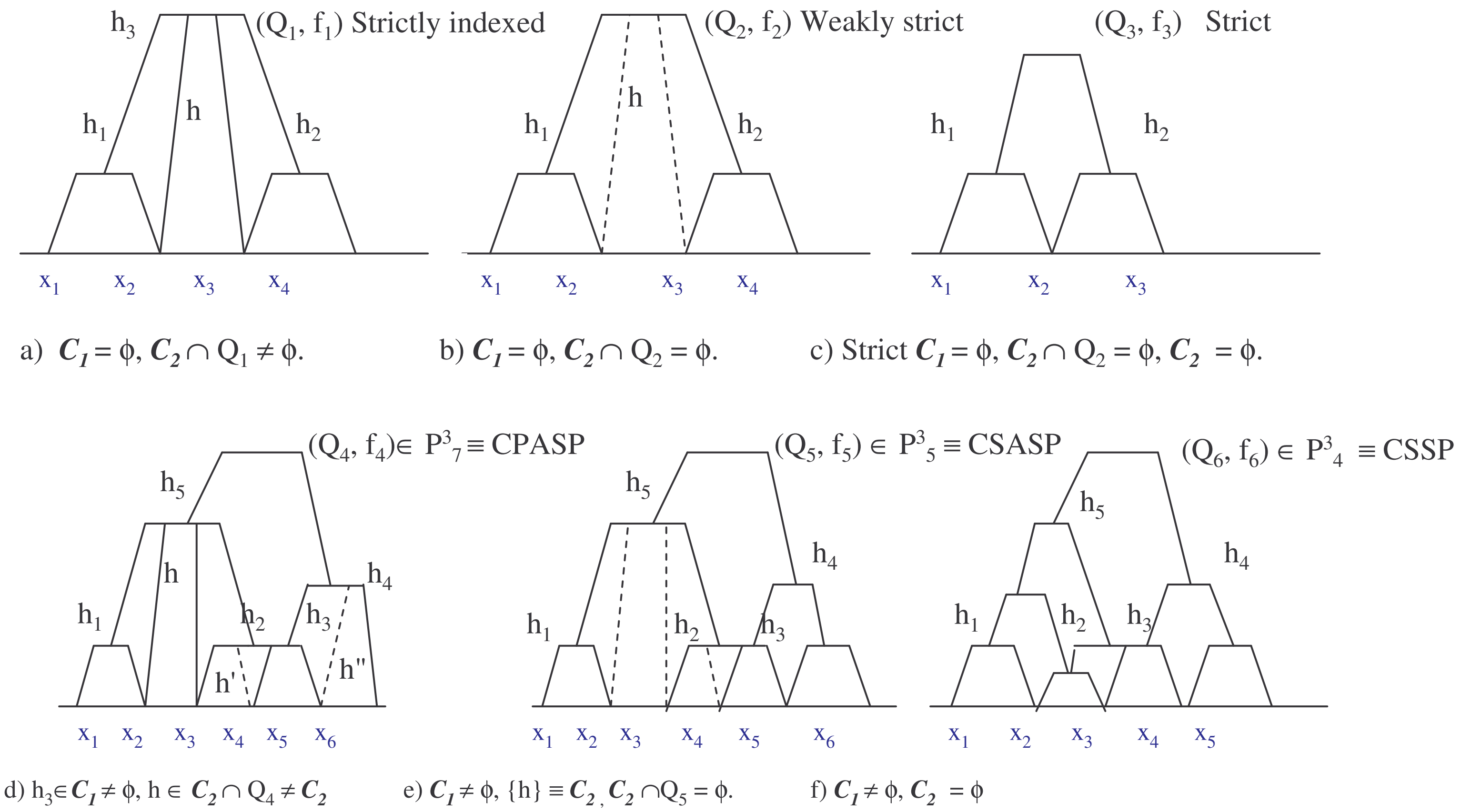
- We say that a spatial pyramid Q (resp a set of indexed convexes of M) is "indexed" by f and (Q, f) is an "indexed spatial pyramid" (resp. a set of indexed convexes of M) if
- f : Q$\rightarrow$ [0, ∞) is such that:
  - $\forall$ A, B $\in$ Q, A $\subset$ B (strict inclusion) $\Rightarrow$ f(A) $\leq$ f(B) ,
  - f(A) = 0 $\Leftrightarrow$ $|A| = 1$.

Three kinds of convex included in a pyramid Q

- ***C*** = set of convexes of the grid M strictly included in an element of Q and with same level.
- ***C****1* = the set of elements C of ***C*** which are the intersection of at least two elements of Q different from C
- ***C****2* = are the other elements of ***C***.

Now, we can define several kinds of indexed spatial pyramids.

Six examples of indexed pyramids



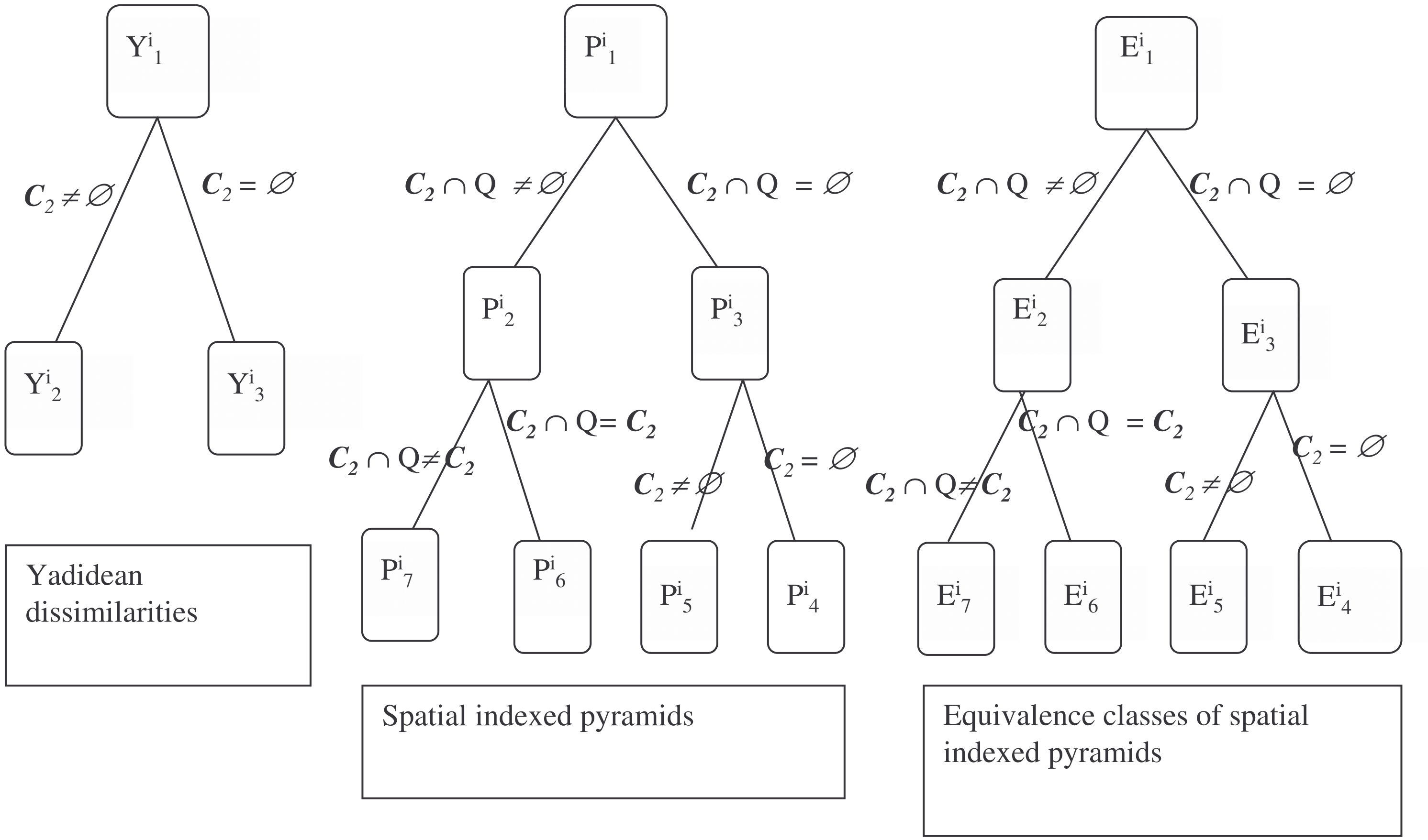
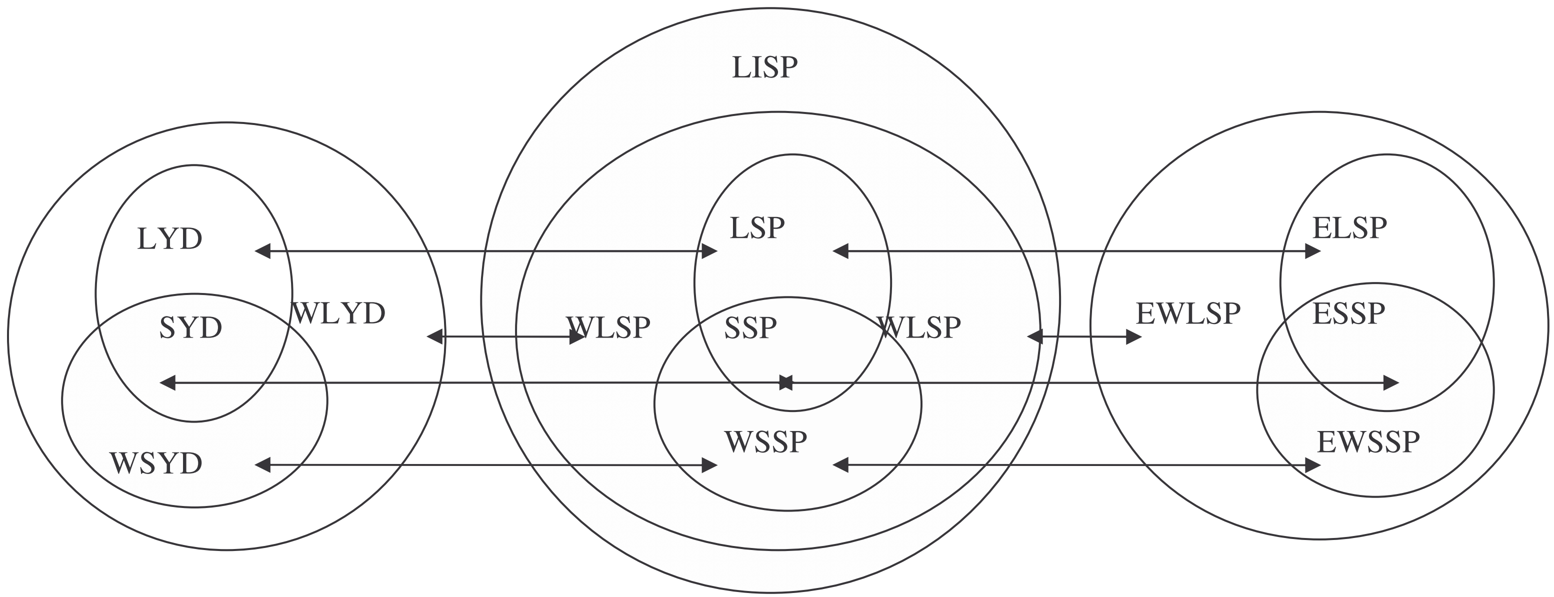


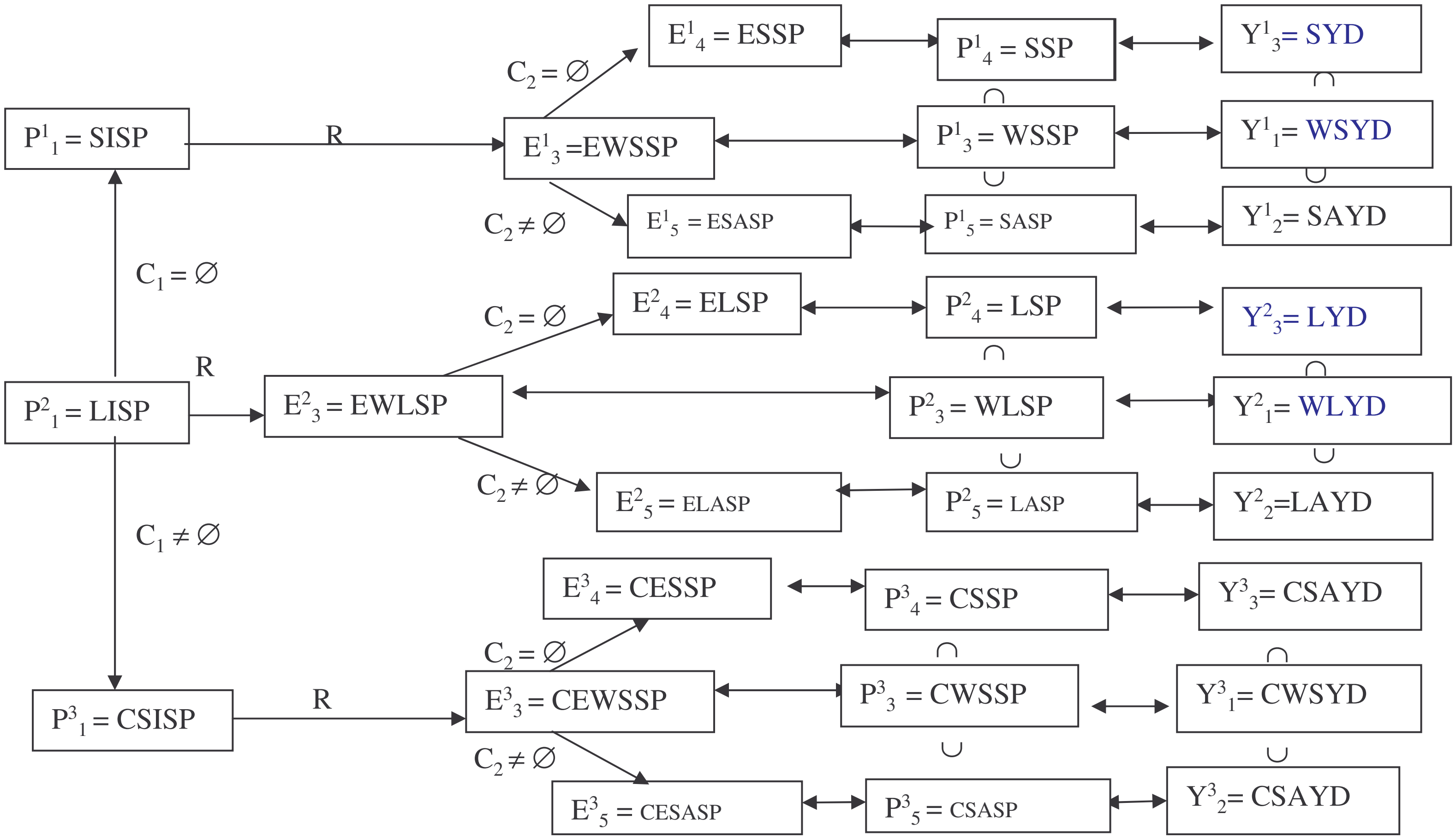
Figure 11: The different cases of Yadidean dissimilarities , Spatial indexed pyramids and Equivalence classes of spatial indexed pyramids. We use the index i such that i = 1 when $\boldsymbol{C}_1$ is empty, i = 2 when $\boldsymbol{C}_1$ may be empty or not empty, i = 3 when $\boldsymbol{C}_1$ is not empty.

Theorem

The set of indexed convex pyramids is in a one-to-one correspondence with the set of Yadidean dissimilarities. This one-to-one correspondence is defined by φ or ψ and moreover $\varphi = \psi^{-1}$, $\psi = \varphi^{-1}$

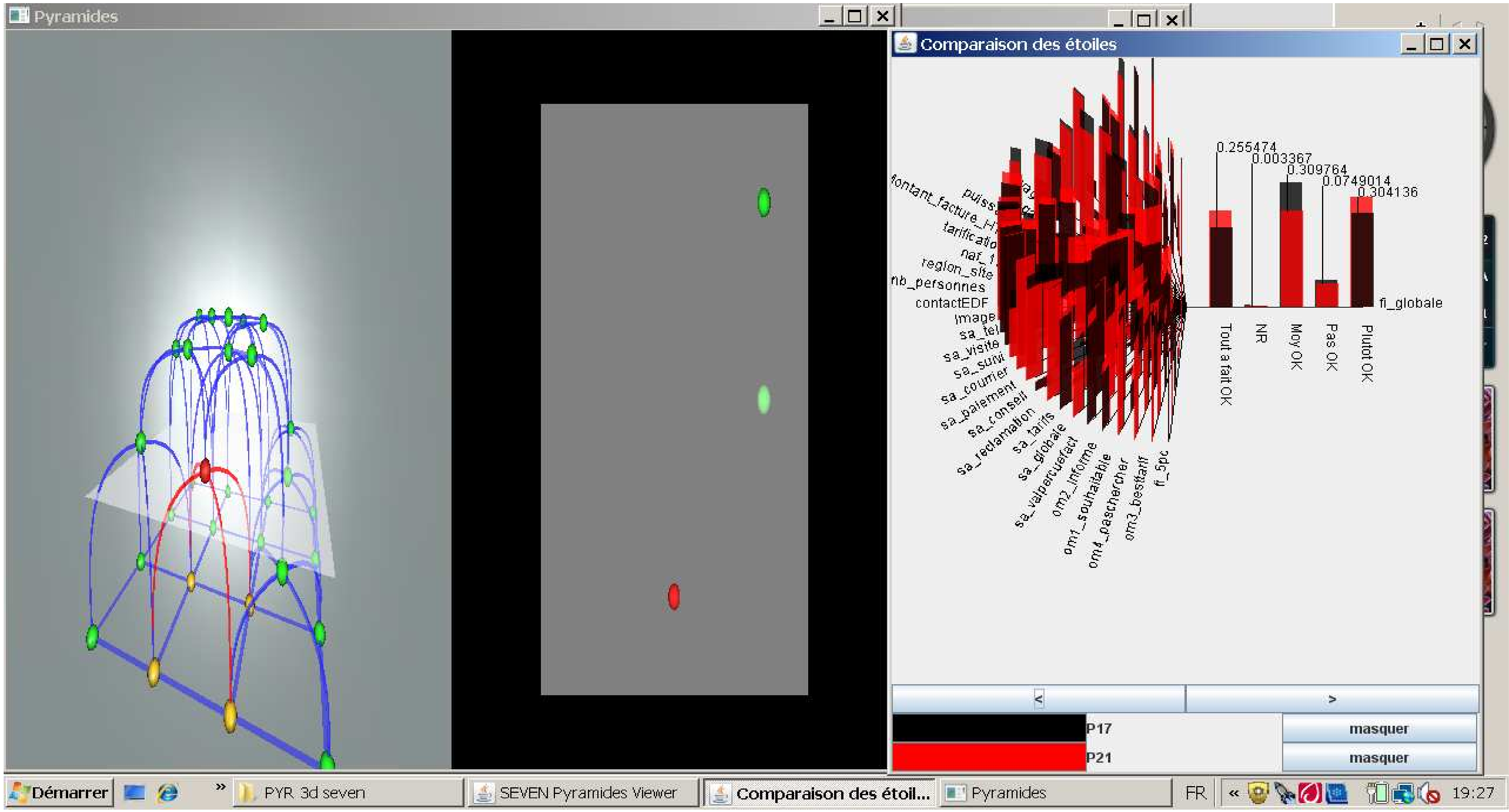
Inclusions and one to one correspondences between Yadidean dissimilarities, indexed spatial pyramids and equivalence classes of spatial pyramids





The main one to one correspondences between indexed spatial pyramids, Yadidean dissimilarities and equivalence classes. Here, 9 one to one correspondences between Yadidean dissimilarities and indexed spatial pyramids are shown among 12 as three more can be added between P^i_6 and Y^i_2 for i = 1, 2, 3.

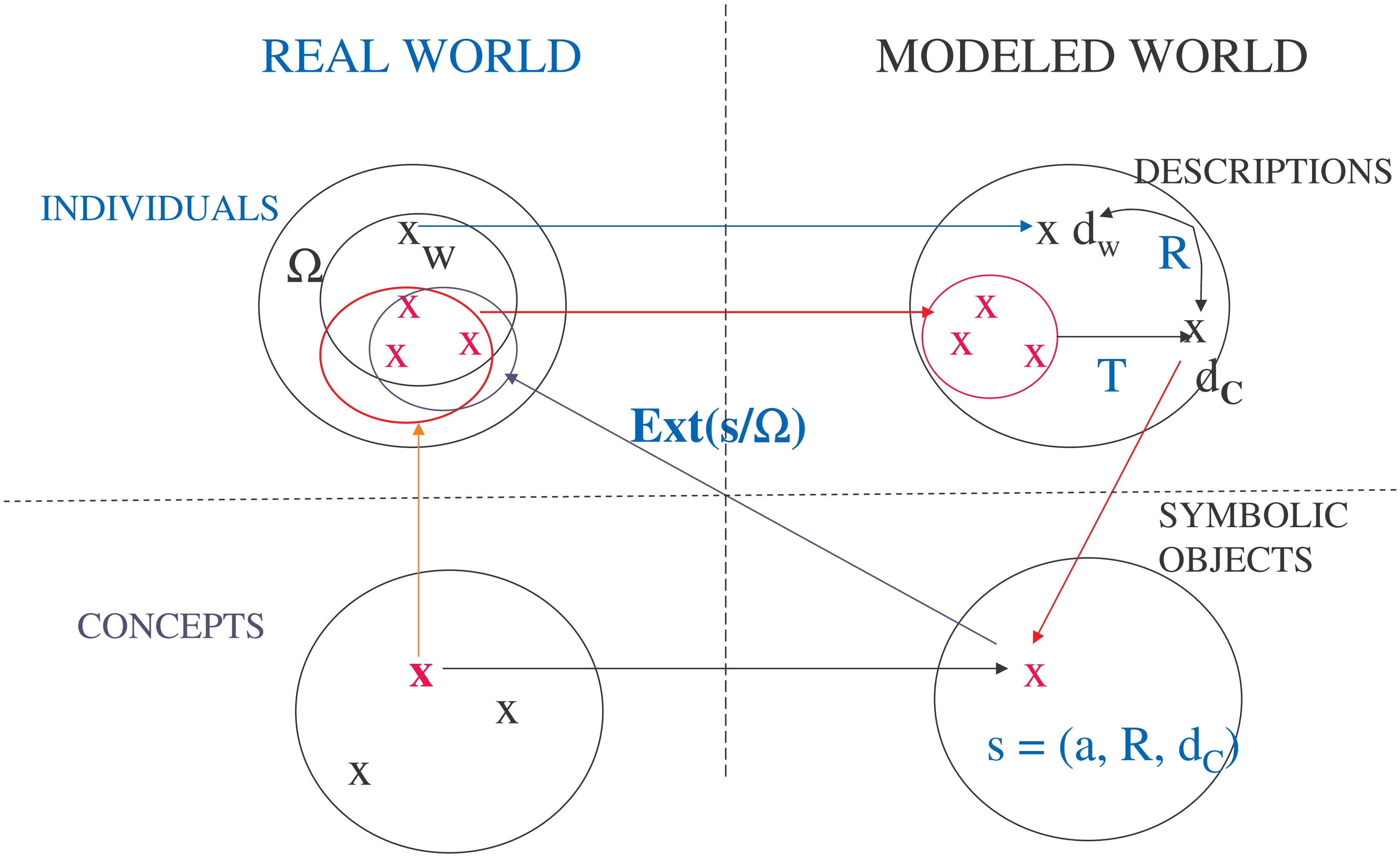
Spatial Pyramidal Software



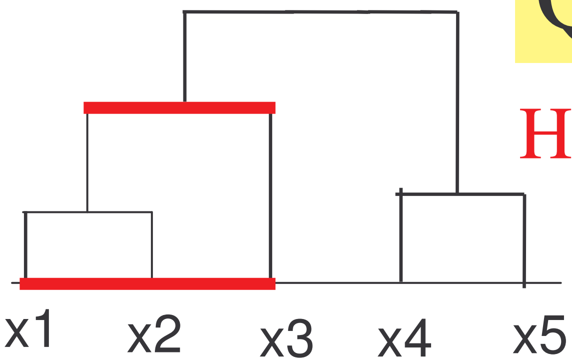
Réalisé dans le cadre de l'ANR SEVEN (EDF, LIMSI, Dauphine).

Théorie de la classification spatiale: E. Diday (2008) "Spatial classification". DAM (Discrete Applied Mathematics) Volume 156, Issue 8, Pages 1271-1294.

QUALITY CONTROL
CONFIRMATORY SDA



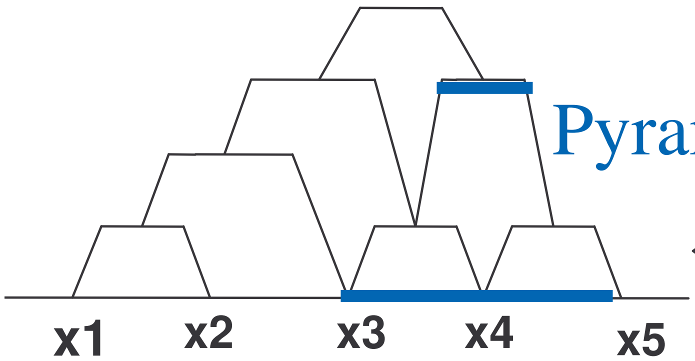
QUALITY CONTROL



Hierarchies

Ultrametric dissimilarity = U

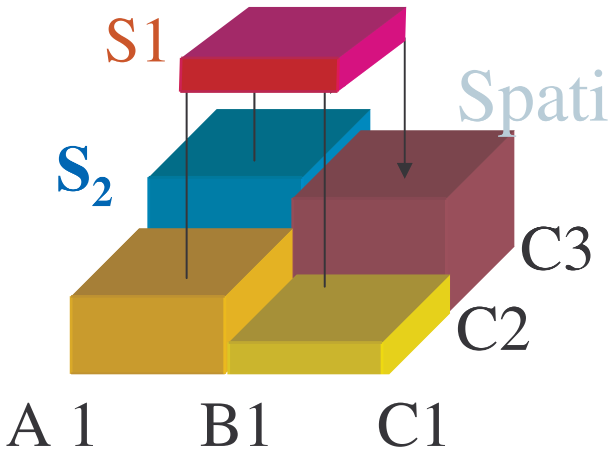
W = |d - U |



Pyramid

Robinsonian dissimilarity = R

W = |d - R |



Spatial Pyramid

Yadidean dissimilarity = Y

W = |d - Y |

CONCLUSION

SYMBOLIC DATA ANALYSIS
allows an extension of learning and exploratory daa analysis to concepts described by data taking care of their internal variation.

It is not better than standard approaches but complementary.

SPATIAL PYRAMIDS
give geometric conceptual structured clusters

reduce distortion with the initial dissimilarity

from standard or symbolic data as input.

much remains to be done:

-a complement for Kohonen maps,

-consensus between spatial pyramids

-by using a volumetric infinite or finite (like a tore) grid, a spatial pyramid can organize and models classes or concepts in a three dimensional space representation.

SYMBOLIC DATA ANALYSIS SOFTWARES

- SODAS (2003) academic from 2 European consortium
- SYR (2008) professional from SYROKKO company

Some References on Spatial Classification

- E. Diday (2004) "Spatial Pyramidal Clustering Based on a Tessellation”. Proceedings IFCS’2004, In Banks et al (Eds.): Data Analysis, Classification and Clustering Methods Heidelberg, Springer-Verlag Springer Verlag.
- E. Diday (2008) “Spatial classification”. DAM (Discrete Applied Mathematics) Volume 156, Issue 8, Pages 1271-1294.
- K. Pak (2005) “ Classifications Hiérarchiques et Pyramidales Spatiales et nouvelles techniques d'interprétation “ Thèse Université Paris Dauphine, 75016 Paris. France.

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.