

Symbolic Data Analysis Of Complex Data:

Edwin Diday
CEREMADE Paris Dauphine University

OUTLINE

- What are Complex Data?
- What are “symbolic data”?
- Why and how symbolic data are built?
- Symbolic Data are Complex Data?
- From Complex Data to Symbolic Data
- What is “Symbolic Data Analysis” (SDA)?
- Tools of SDA
- Some industrial applications:
Nuclear Power Plot, Text Mining, TGV on a bridge,
Funds
- Open directions of research
- Conclusion: SDA provides a framework for Complex Data Analysis (CDA)

OUTLINE

- **What are Complex Data?**
- What are “symbolic data”?
- How “Symbolic Data” are built?
- 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 built?
- 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), ...}

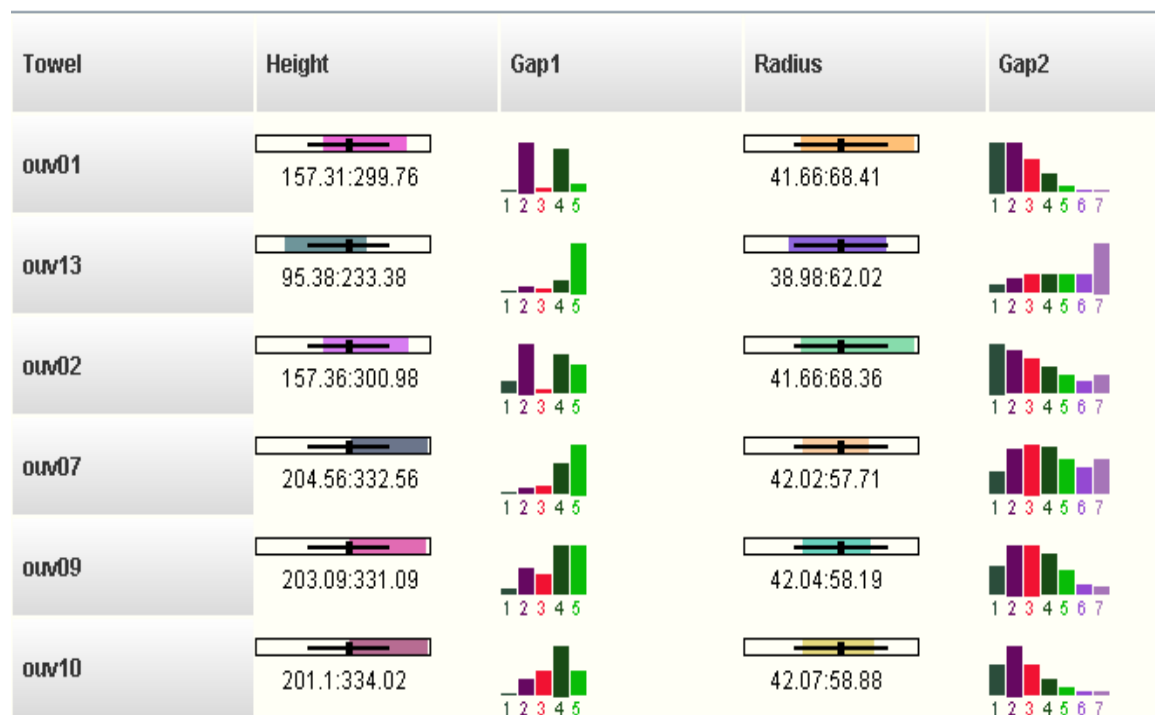
Here the variation (of weight, nationality, ...) concerns the players of each team.

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.**

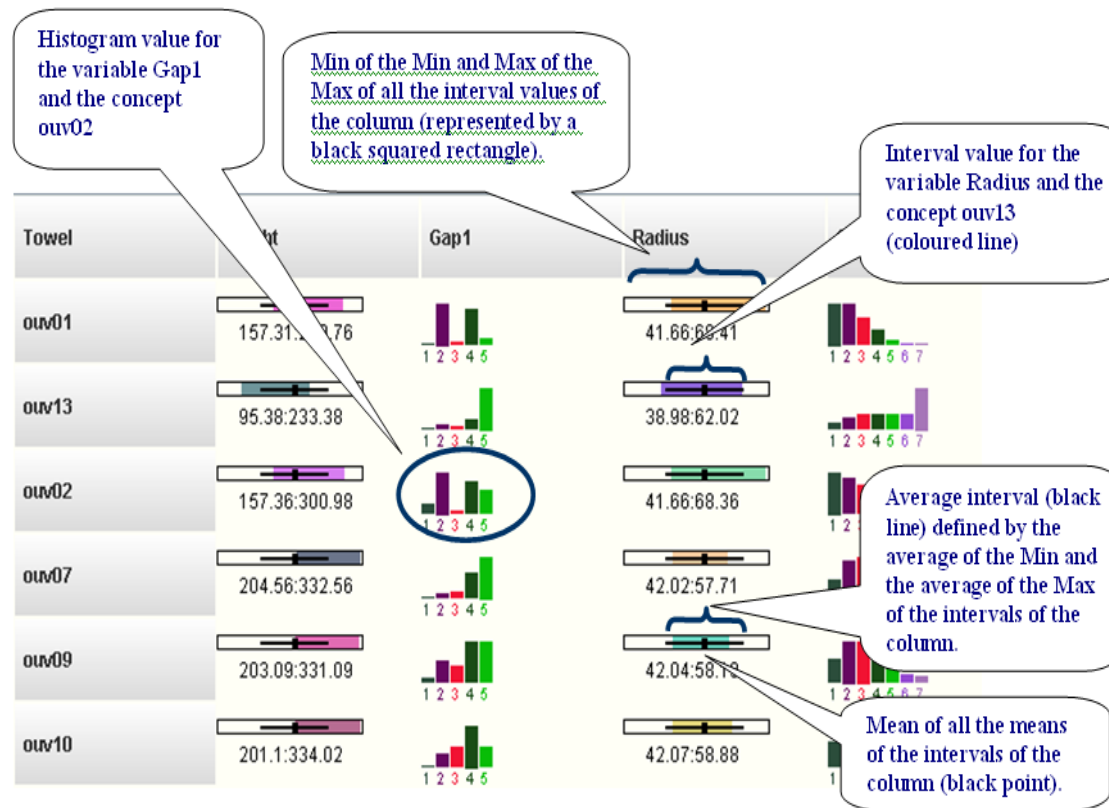
SYMBOLIC DATA TABLE SOFTWARE*



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 afonso@syrokko.com

OUTLINE

- What are Complex Data?
- What are “symbolic data”?
- **When and how symbolic data are built?**
- 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.

WHEN SYMBOLIC DATA ANALYSIS?

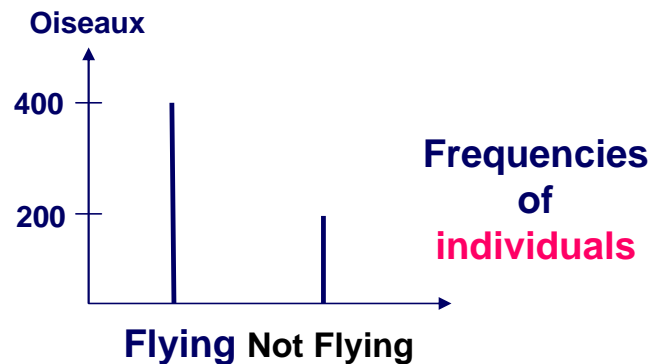
- When the good units are the concepts: finding why a team is a winner is not finding why a player is a winner
- When the categories of the class variable to explain are considered as new units and described by explanatory symbolic variables.
- When the initial data are composed by multisource data tables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

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	Fly	Color	Size (cm)
1	penguins	No	black	80
2	swallows	yes	grey	30
600	ostriches	No	black	125

swallows, ostriches, and penguins are the “concepts”



Symbolic Data Table

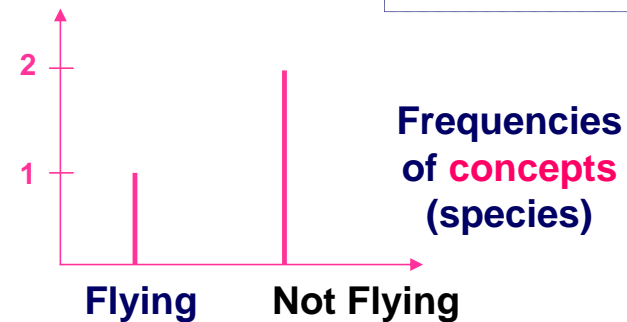
Species	Fly	Color	Size	Migr
swallows	yes	0.3b,0.7grey	[25, 35]	Yes
ostriche	No	0.1black,0.9g	[85,160]	No
Penguin	No	0.5b,0.5grey	[70, 95]	Yes

The species are the new units

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

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

Species



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

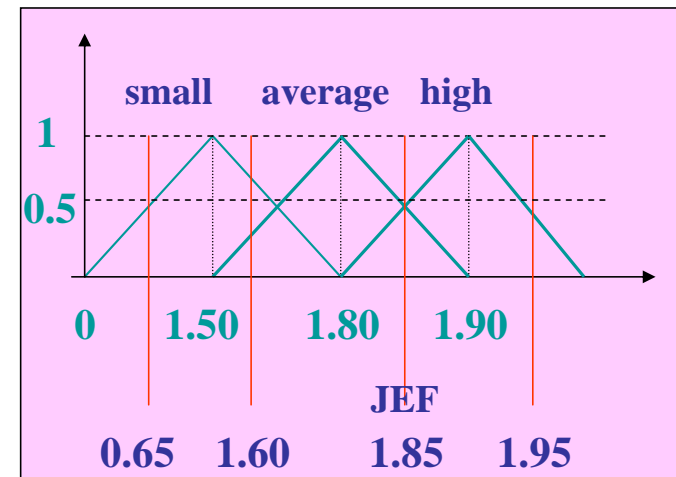
	height			weight	hair
	small	average	high		
Paul	0.70	0.30	0	45	yellow
Jef	0	0.50	0.50	80	yellow
Jim	0.50	0	0	30	black
Bill	0	0	0.48	90	black

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

From Numerical to Fuzzy Data



When? Example of 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%)}}	

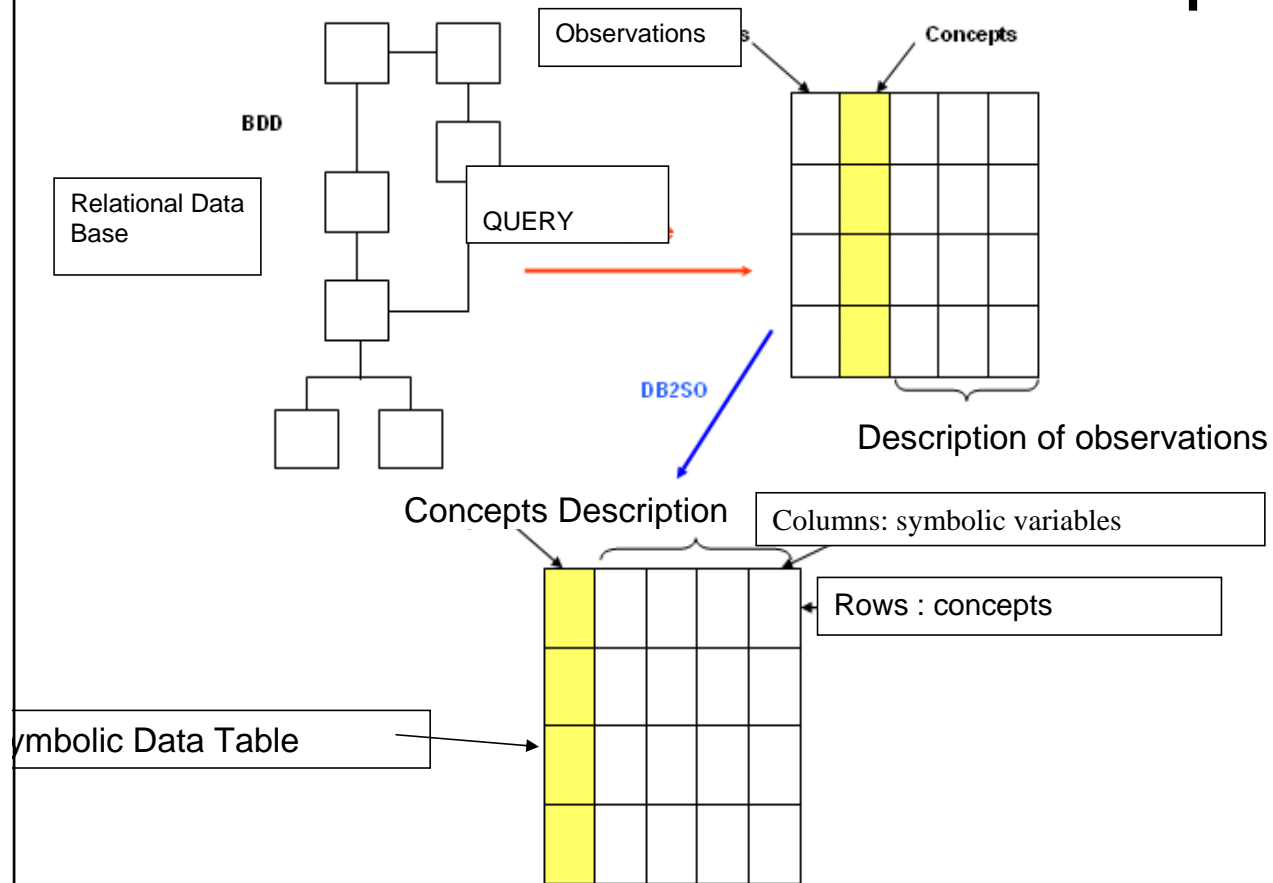


Concatenation

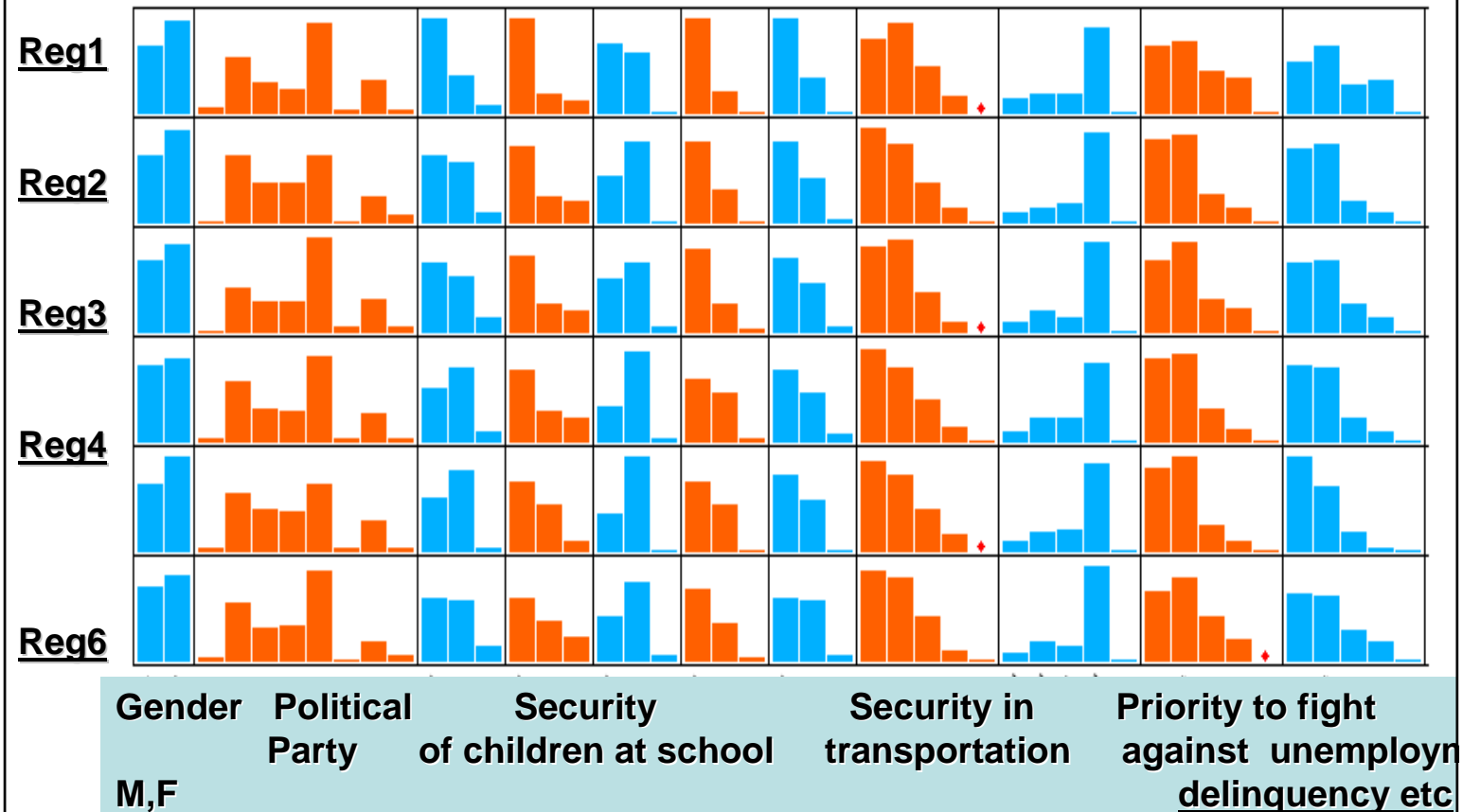
IRIS n = [Symb. Description of households] \wedge [Symb. Description of School]
NEW DATA: in one SYMBOLIC DATA TABLE describing each IRIS.

HOW?

From Database to Concepts



Tackle security problems in regions



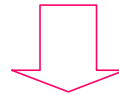
OUTLINE

- What are Complex Data?
- What are “symbolic data”?
- Why and how symbolic data are built?
- **Symbolic Data are Complex Data?**
- From Complex Data to Symbolic Data
- What is “Symbolic Data Analysis” (SDA)?
- Open directions of research
- Conclusion: SDA provides a framework for Complex Data Analysis (CDA)

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

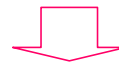
Symbolic Data Table

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	Weight Max	Size Min	Size Max	Eur	Afr
Very good	80	95	1.70	1.95	0.7	0.3

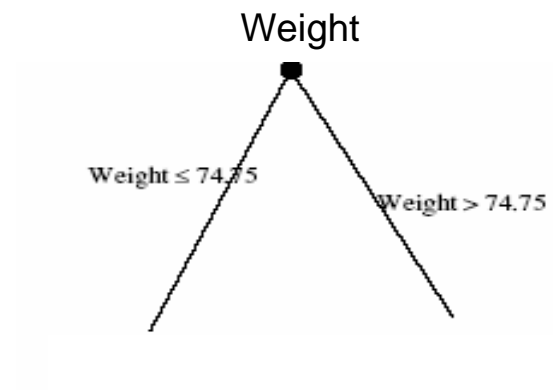


Concern:

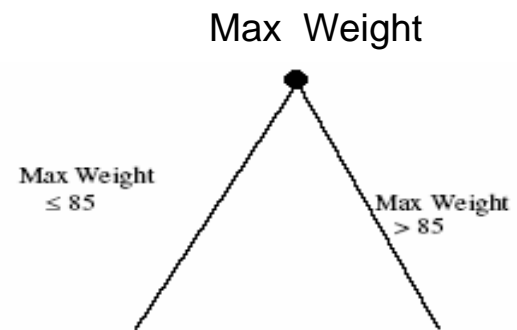
The initial variables are lost and the variation is lost!

Divisive Clustering or Decision tree

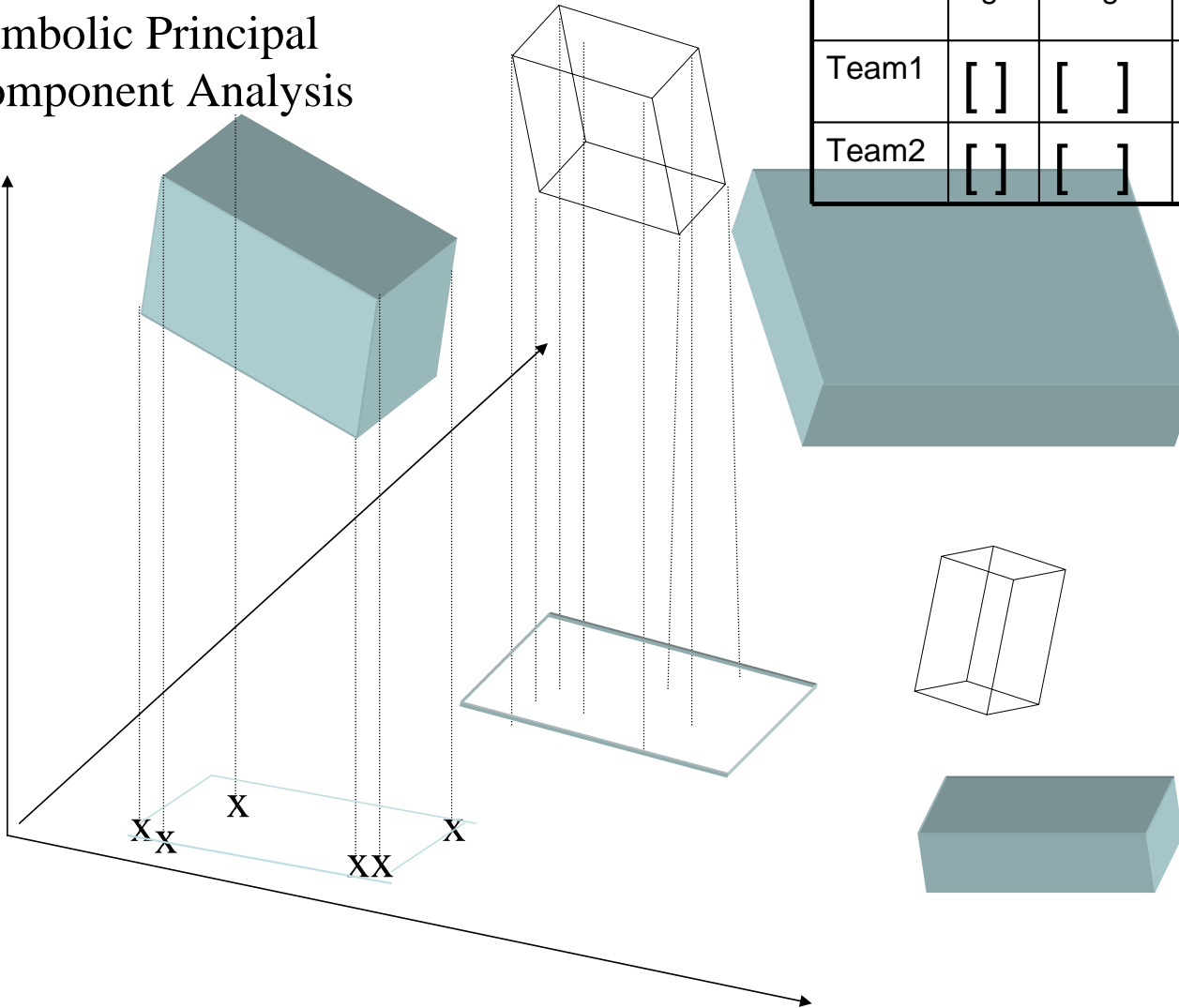
Symbolic Analysis



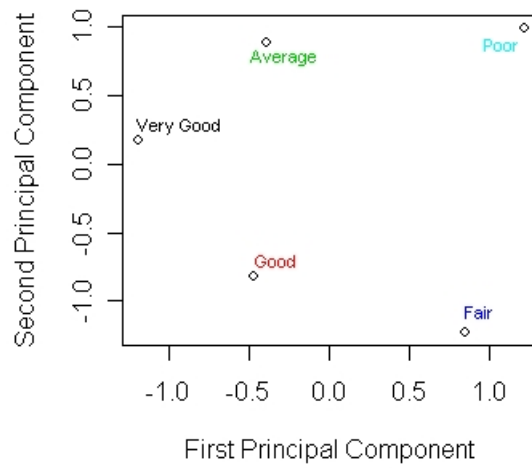
Classical Analysis



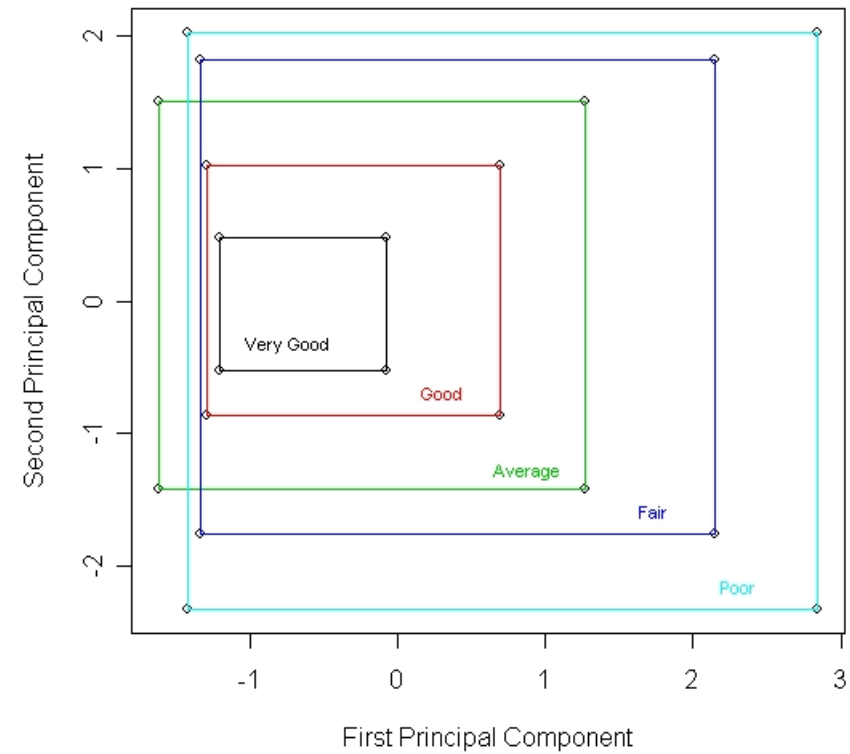
Symbolic Principal Component Analysis



	age	weight	height
Team1	[]	[]	[]
Team2	[]	[]	[]

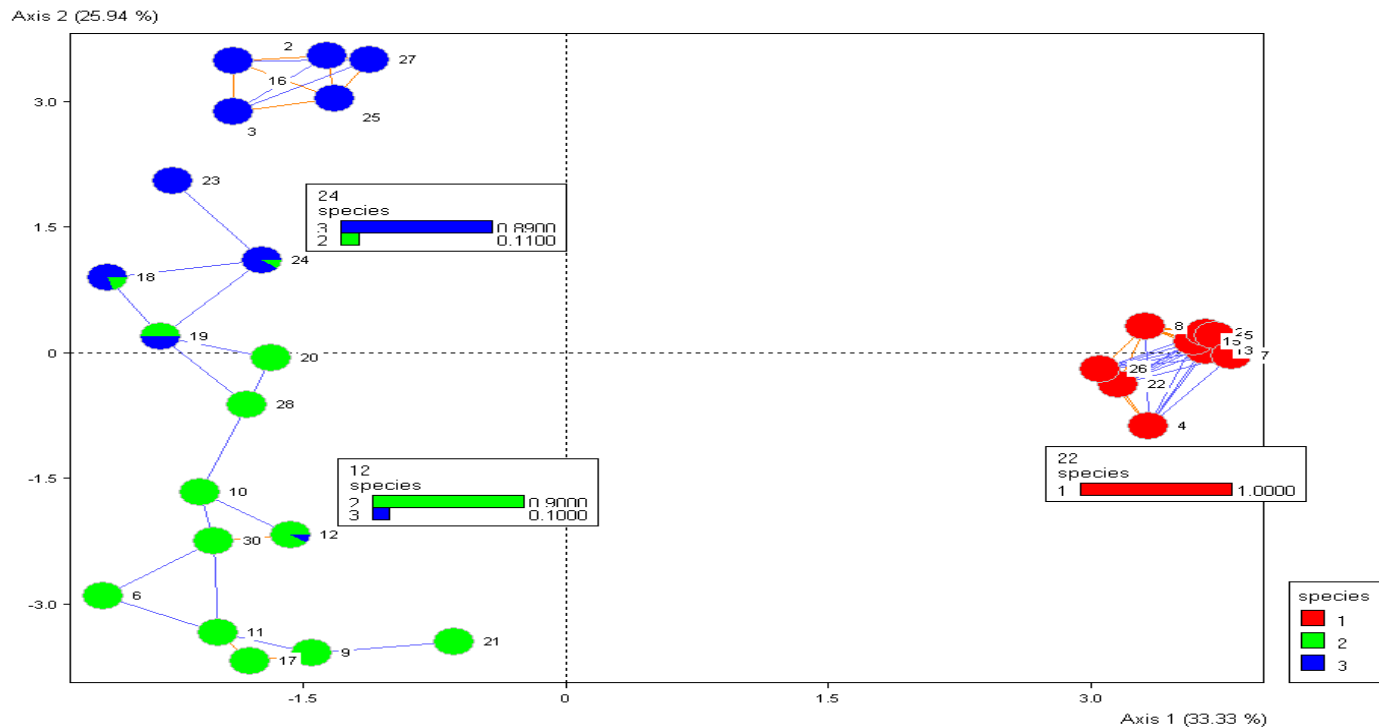


Classical Analysis
Loose variation



Symbolic Analysis
Take care of
variation

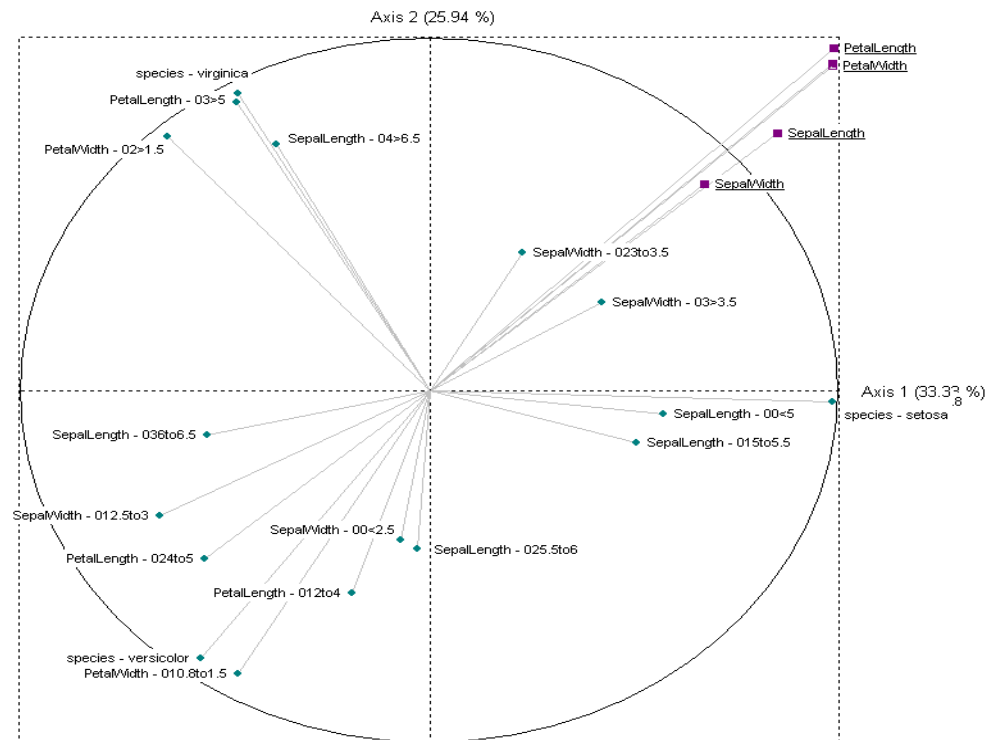
PCA and NETWORK OF BAR CHART DATA of 30 Iris Fisher Data Clusters*



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

* SYROKKO Company afonso@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 losing much information.

OUTLINE

- What are Complex Data?
- What are “symbolic data”?
- How “Symbolic Data” are built?
- 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

INTERVAL TIME SERIES VOLATILITY OF STOCKS

The symbolic aggregation approach

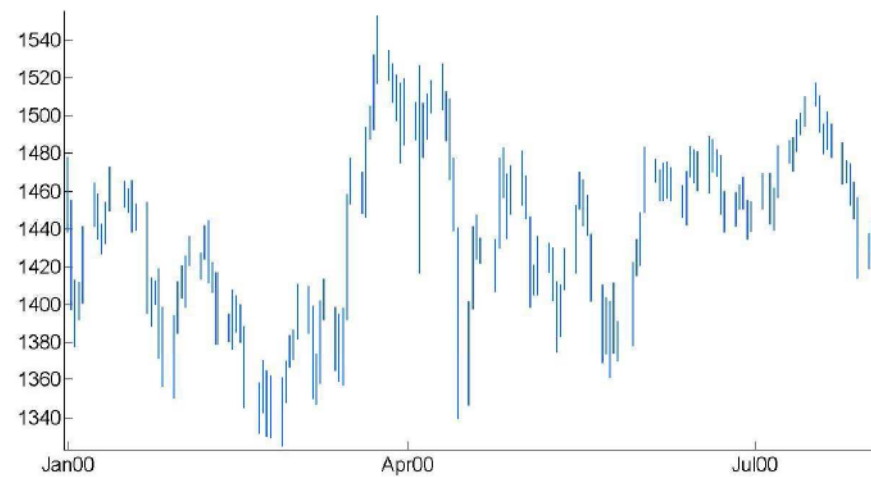
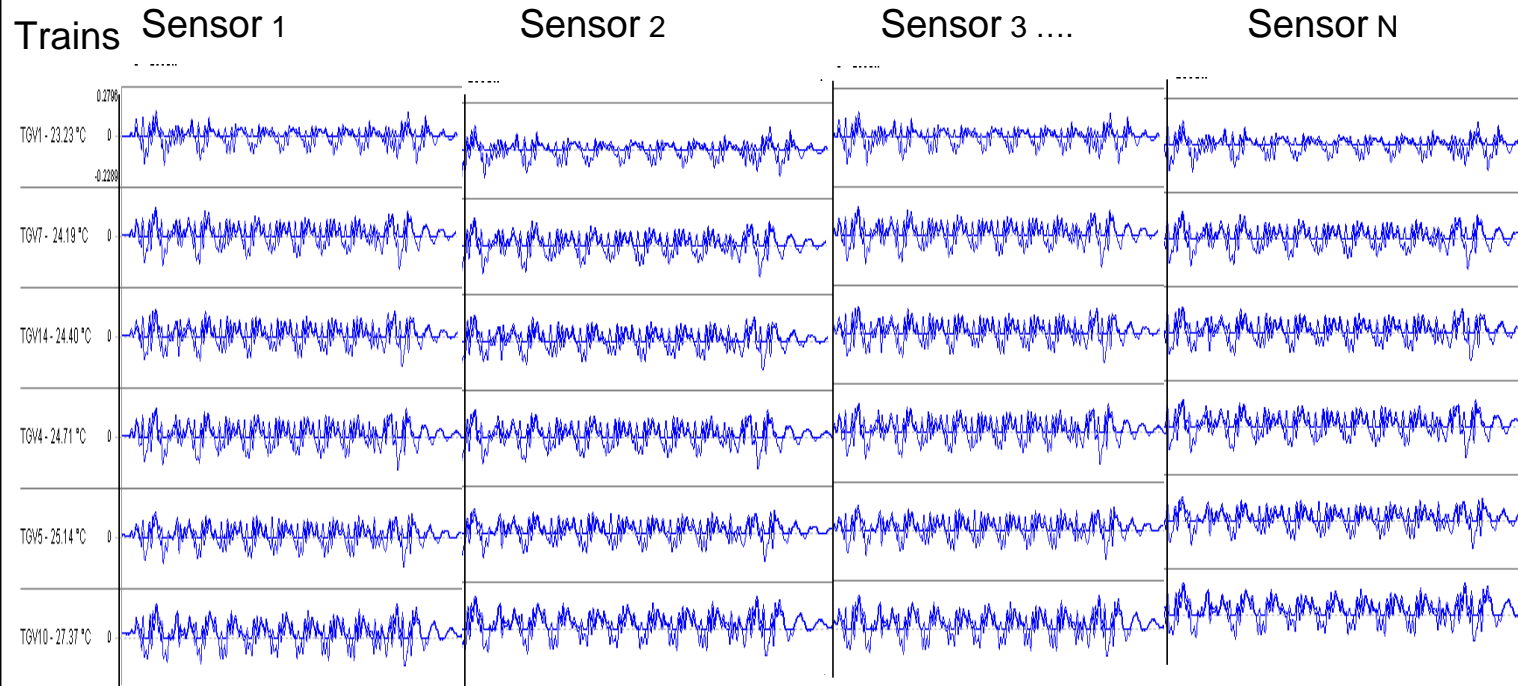


Figure 2: Interval time series of the high and low prices of the SP500 stock index.

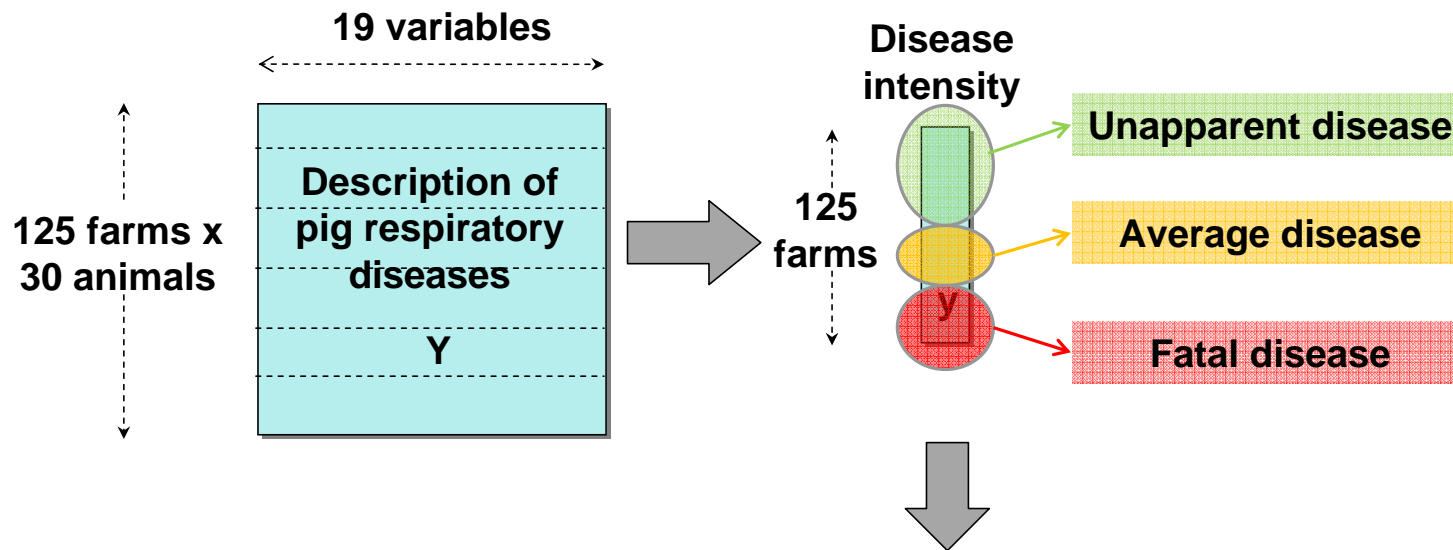
Time Series Data table: Anomaly detection on a bridge LCPC (Laboratoire Central Des Ponts et Chaussées) and SNCF Data



Each row represents a train going on the bridge at a given temperature,
each cell contains until 800.000 values.

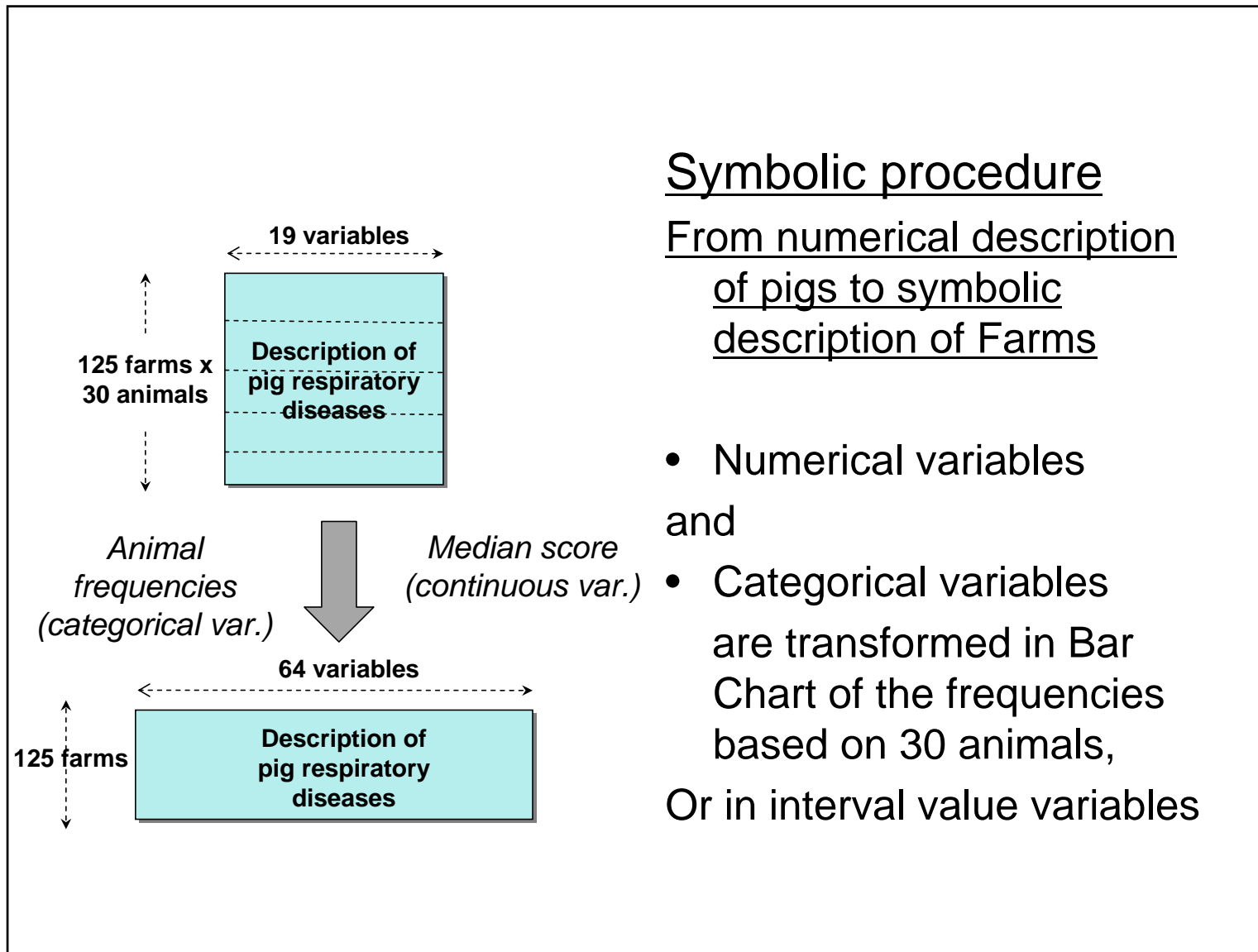
Each cell is transformed in HISTOGRAM from a PROJECTION or from WAVELETS

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 afonso@syrokko.com

OUTLINE

- What are Complex Data?
- What are “symbolic data”?
- Why and how symbolic data are built?
- Symbolic Data are Complex Data?
- From Complex Data to Symbolic Data
- **What is “Symbolic Data Analysis” (SDA)?**
- Tools of SDA
- Some industrial applications:
 - Nuclear Power Plot, Text Mining
- Open directions of research
- Conclusion: SDA provides a framework for Complex Data Analysis (CDA)

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

TO

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

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!**

THE TWO STEPS OF A SDA

- 1) Building the symbolic data from the Data Base
- 2) Applying SDA methods.

OUTLINE

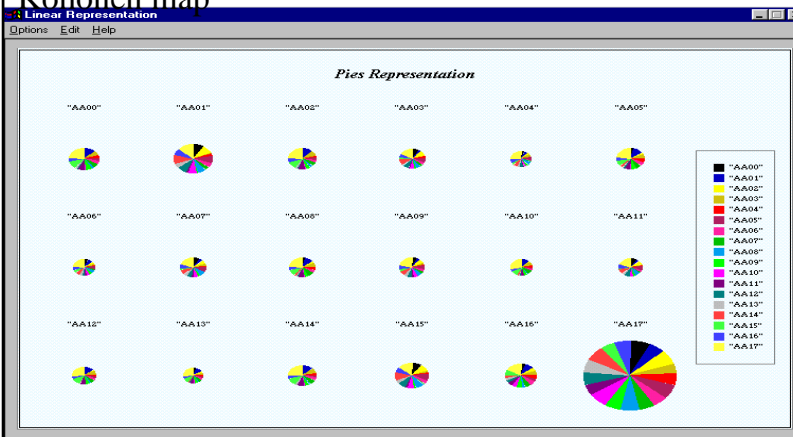
- What are Complex Data?
- What are “symbolic data”?
- Why and how symbolic data are built?
- Symbolic Data are Complex Data?
- From Complex Data to Symbolic Data
- What is “Symbolic Data Analysis” (SDA)?
- **Tools of SDA**
- Open directions of research
- Conclusion: SDA provides a framework for Complex Data Analysis (CDA)

SYMBOLIC DATA ANALYSIS TOOLS HAVE BEEN DEVELOPPED

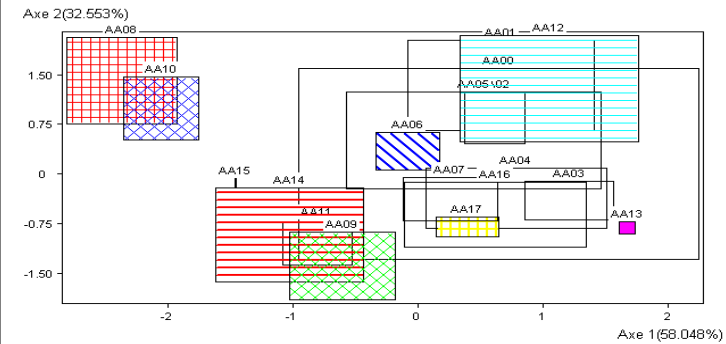
- **Graphical visualisation of Symbolic Data**
- **Correlation, Mean, Mean Square Histogram of a symbolic variable**
- **Dissimilarities between symbolic descriptions**
- **Clustering of symbolic descriptions**
- **S-Kohonen Mappings**
- **S-Decision Trees**
- **S-Principal Component Analysis**
- **S-Discriminant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

Examples of SDA Output of the software

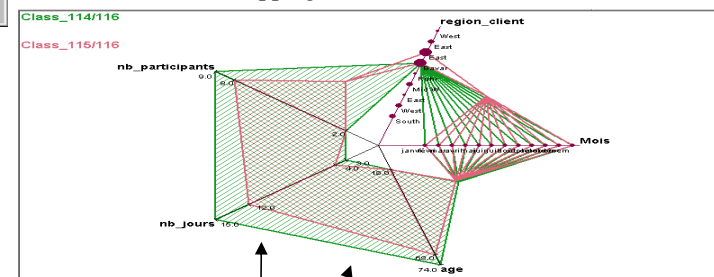
Kohonen map



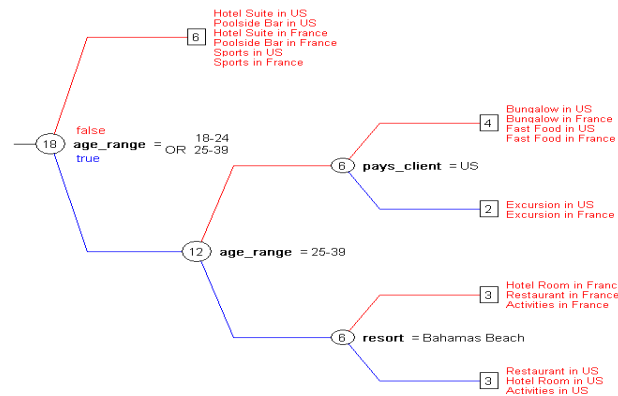
Principal component



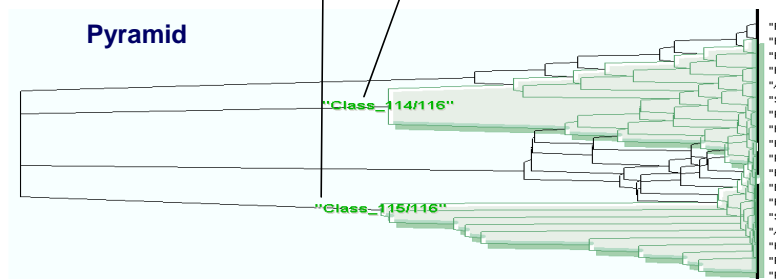
Zoom stars overlapping



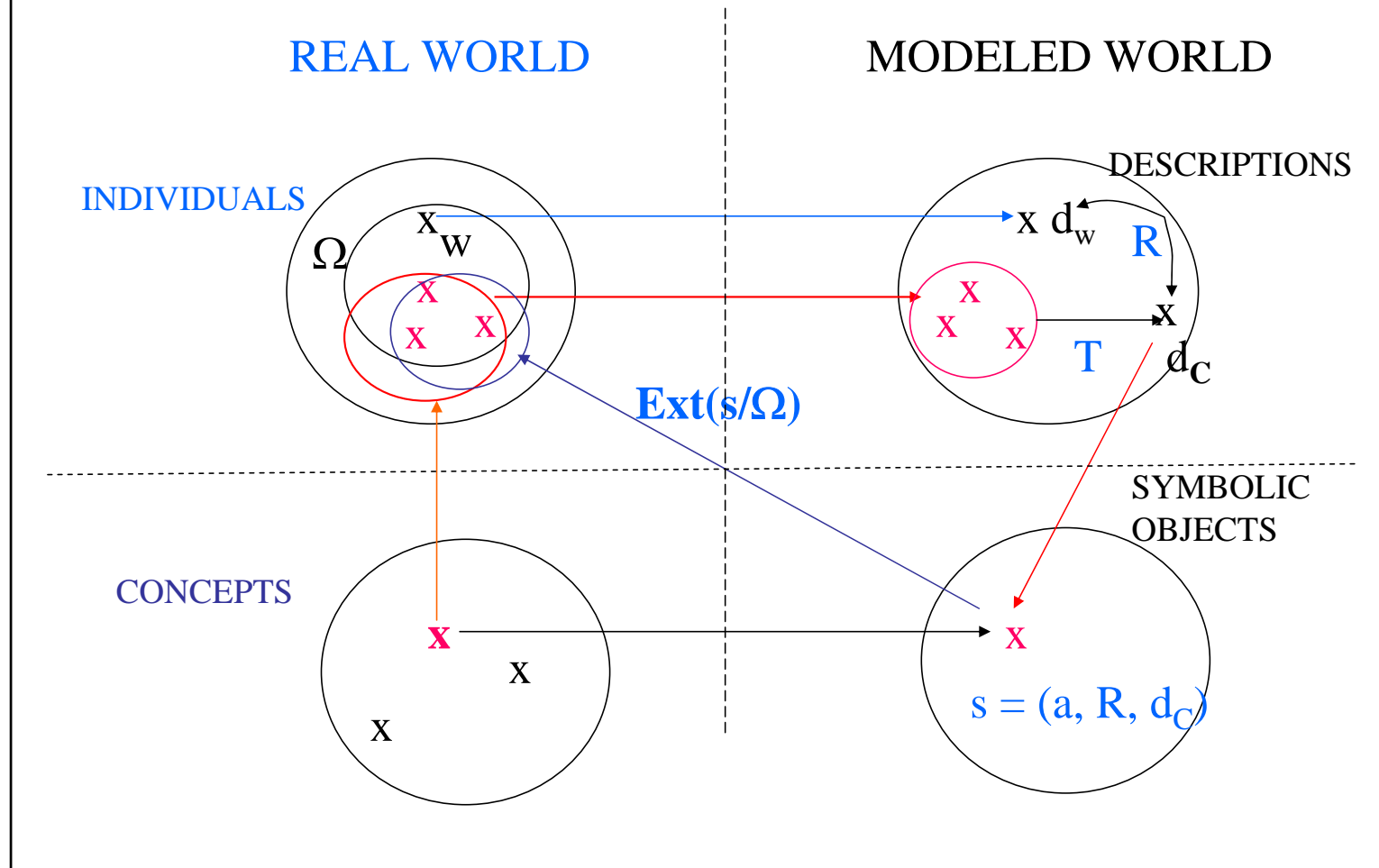
Top down clustering tree



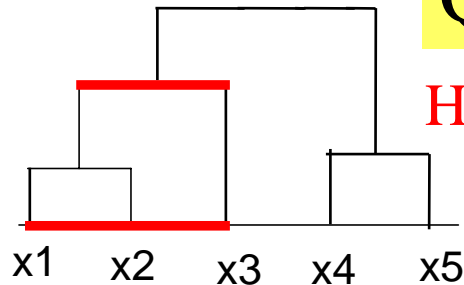
Pyramid



QUALITY CONTROL CONFIRMATORY SDA



QUALITY CONTROL

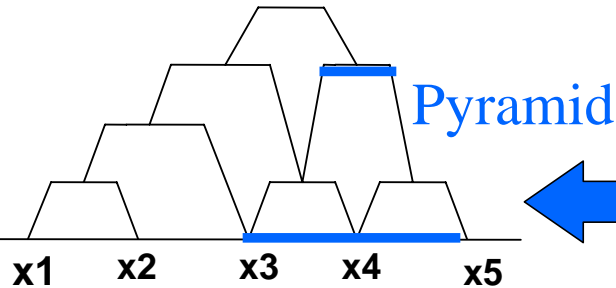


Hierarchies



Ultrametric
dissimilarity = U

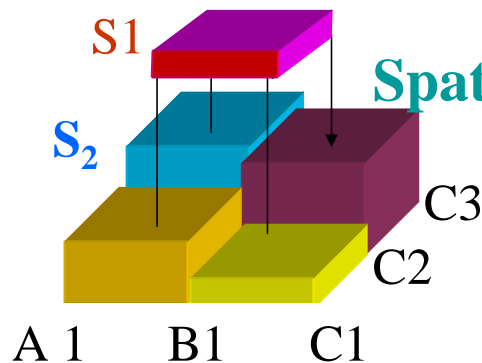
$$W = |d - U|$$



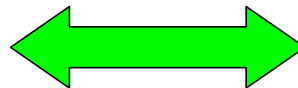
Pyramid



Robinsonian
dissimilarity = R



Spatial Pyramid

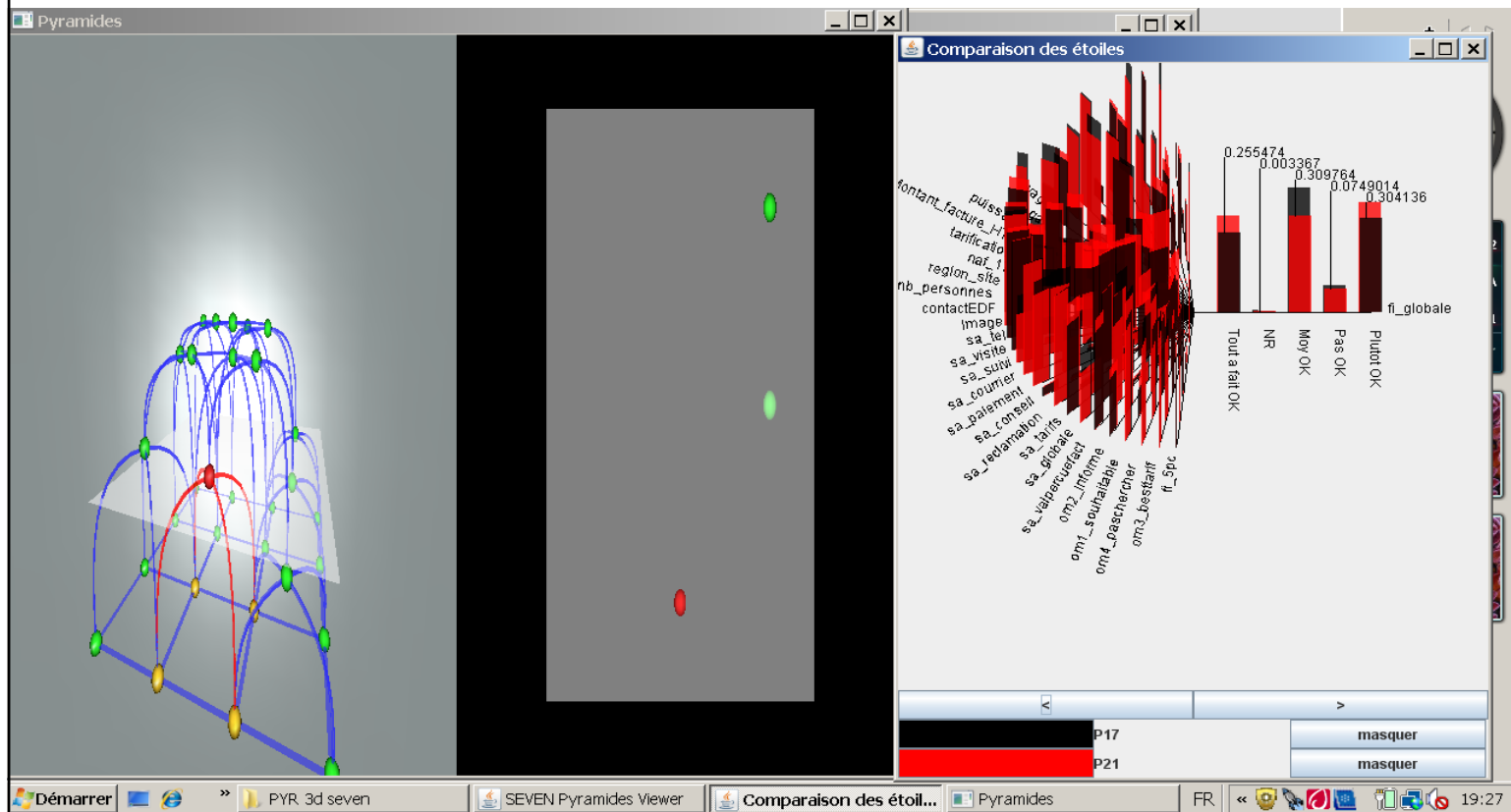


$$W = |d - R|$$

Yadidean
dissimilarity = Y

$$W = |d - Y|$$

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.

OUTLINE

- What are Complex Data?
- What are “symbolic data”?
- Why and how symbolic data are built?
- Symbolic Data are Complex Data?
- From Complex Data to Symbolic Data
- What is “Symbolic Data Analysis” (SDA)?
- Tools of SDA
- **Some industrial applications:**
 - Nuclear Power Plot, Text Mining, TGV anomalies**
- Open directions of research
- Conclusion: SDA provides a framework for Complex Data Analysis (CDA)

Nuclear Power Plant

Find Correlations Between
3 Standard Data Tables of Different
observation units and different
Variables

NUCLEAR POWER PLANT

Nuclear thermal power station

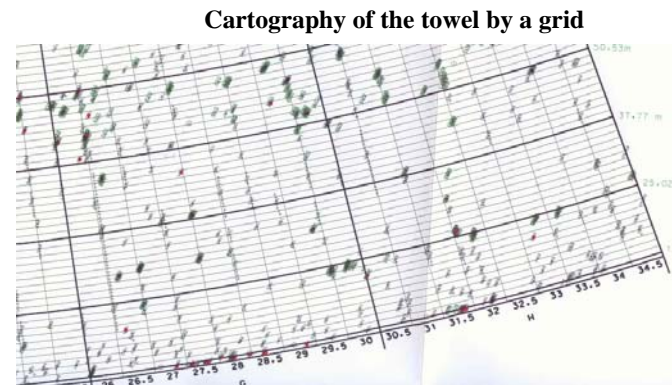
Inspection :



Inspection machine



Craks



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

Telephone calls text mining in order to discover “themes” without using semantic

INITIAL DATA: 2 814 446 rows

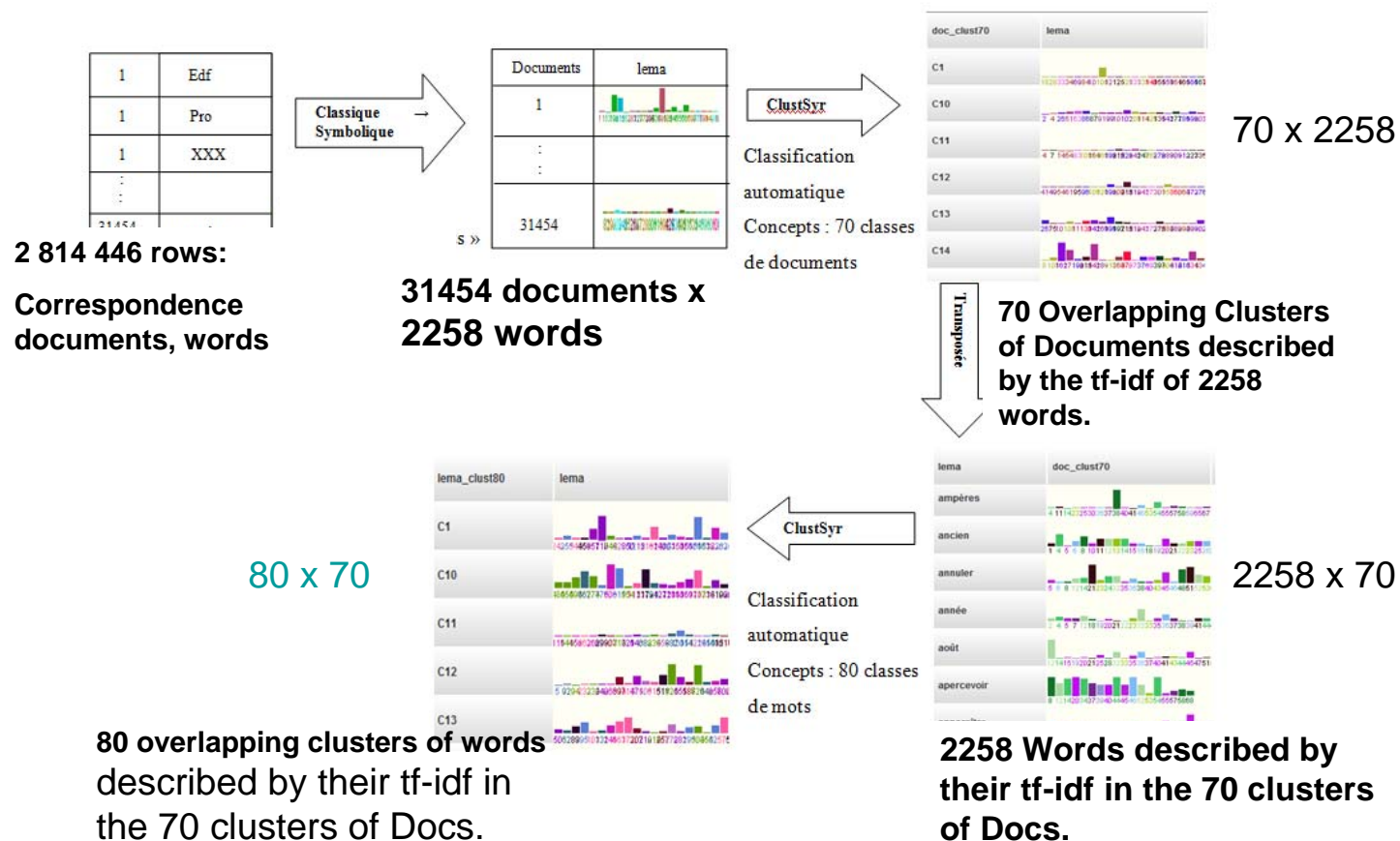
Documents	Words
Doc1	bonjour
Doc1	oui
Doc1	monsieur
.....	
Doc2	panne
.....	

Correspondence between words and documents.

Each calling session is called a document. We start after lemmatisation with a table of

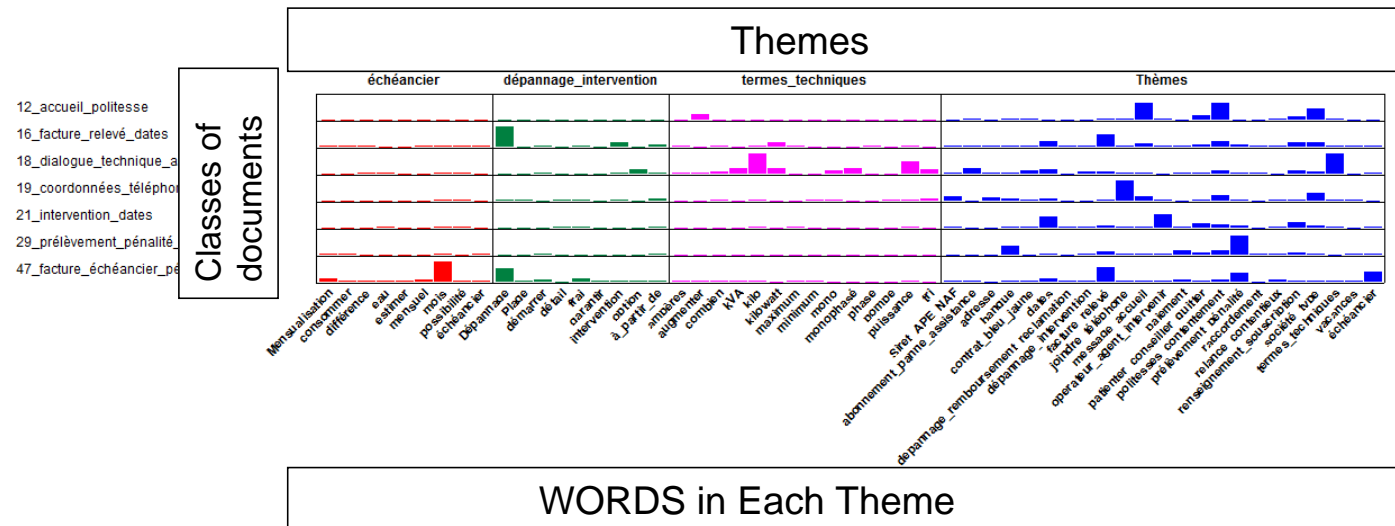
- 31454 documents
- 2258 words

First Steps: building overlapping clusters of documents and words: **CLUSTSYR**



Next step:
STATSYR

Each cluster of documents is described by the 80 clusters of words called
“themes”



NEXT STEPS: SELECTION

- BEST WORDS: HISTSYR
- BEST THEMES: TABSYR
- BEST CLASSES OF DOCUMENTS: TABSYR
- GRAPHICAL REPRESENTATION of themes , doc classes, clusters: NETSYR
- SOCIAL NETWORK: NETSYR
- ANNOTATION of Themes and Document classes : NETSYR

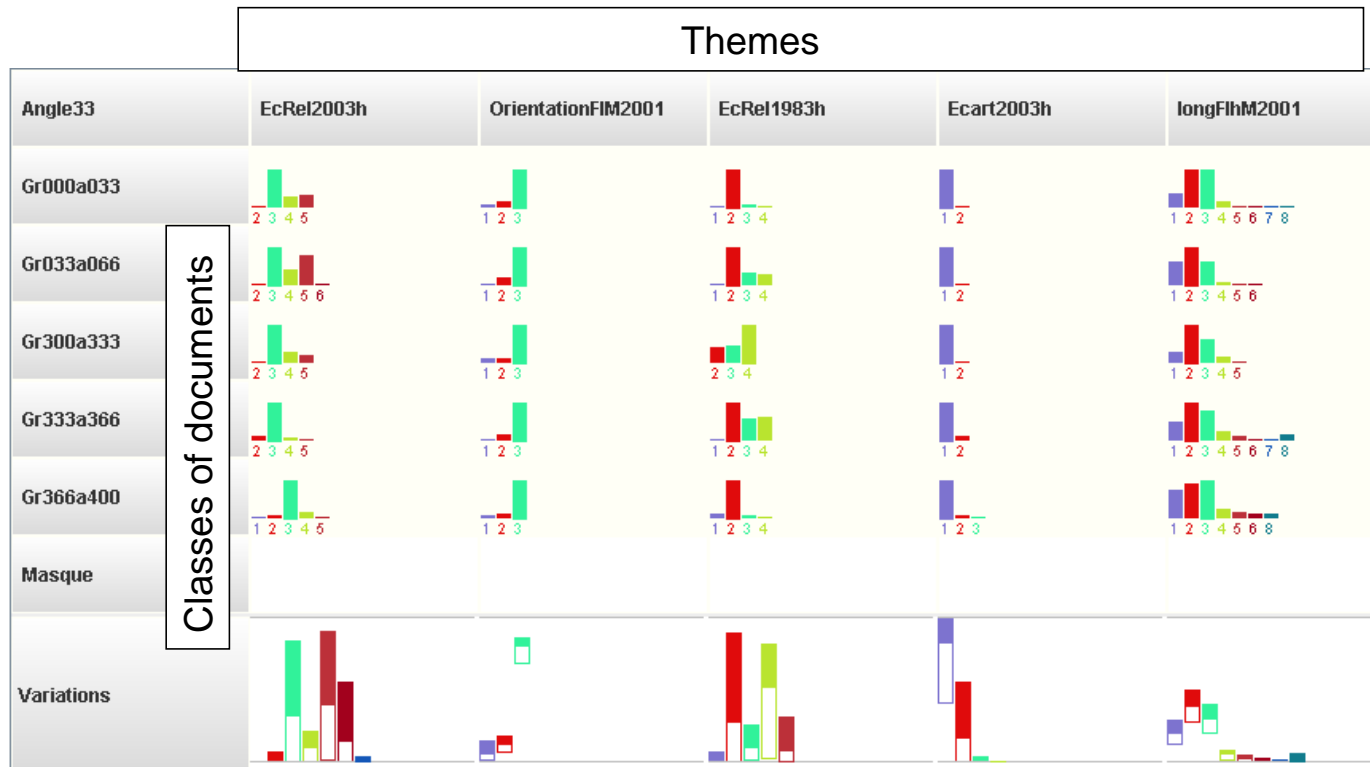
"BEST WORDS" SELECTION

- **HistSyr** allows the selection of the best discriminating words among the document classes: the score is obtained by using the variation of the distributions between the concepts (classes of documents).



BEST THEMES : TABSYR

BEST DOCUMENT CLASSES : TABSYR



SORTING AND SELECTION of THE “BEST THEMES”: the score is obtained by using the variation of the distributions between the concepts.

GRAPHICAL REPRESENTATION

NETSYR

GRAPHICAL REPRESENTATION

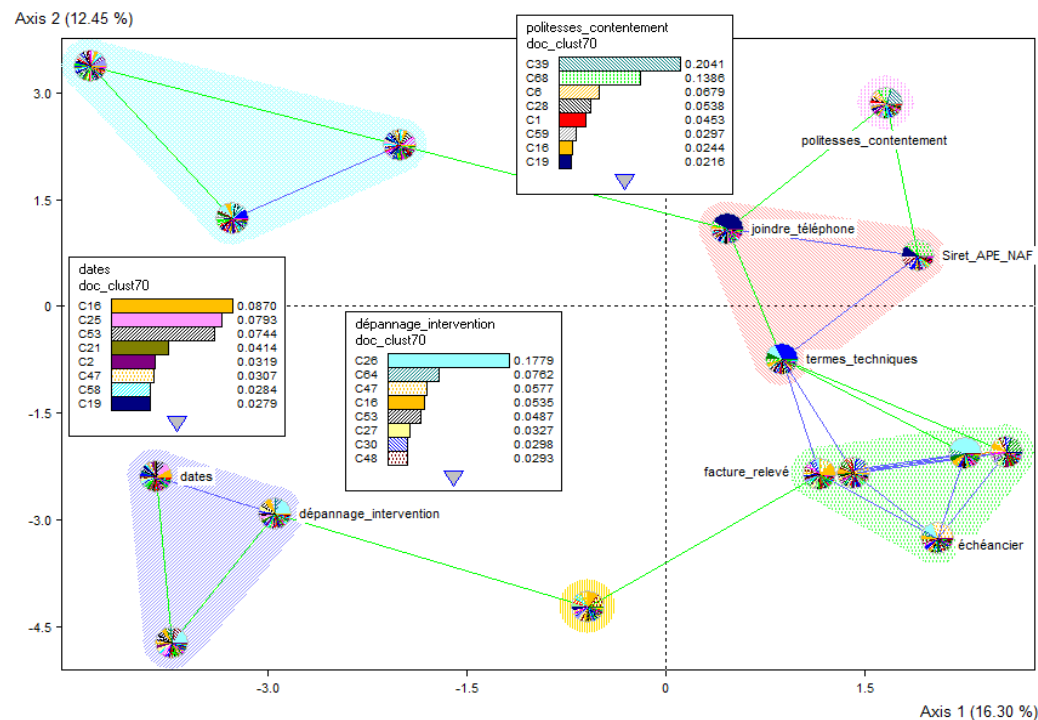
of themes ,
document classes, by
Pie Charts
And their Bar chart
description.

Overlapping Clusters

SOCIAL NETWORK Based on dissimilarities

ANNOTATION :
of Themes and
Document classes

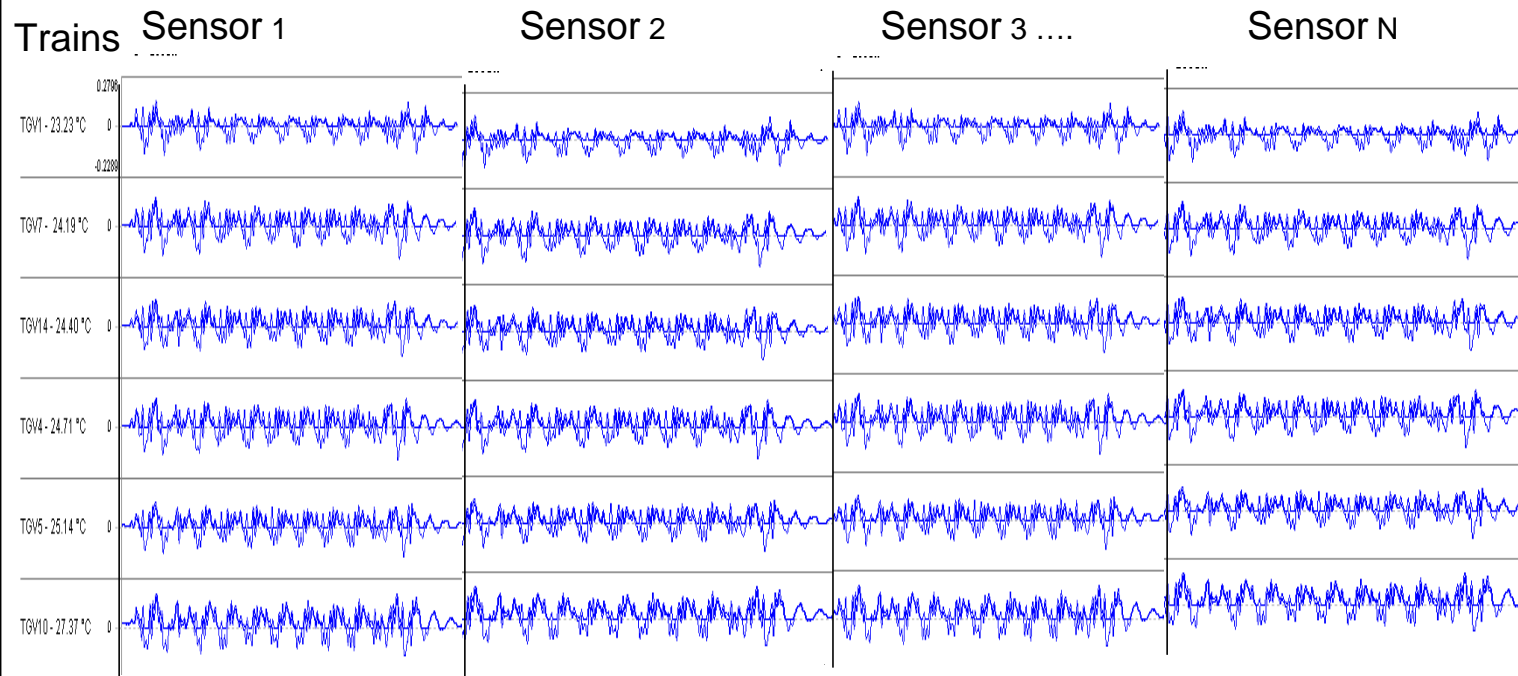
Moving, Zooming...



We obtain finally a clear representation of the main themes , their classes and their links : “failures”, “budget”, “addresses”, “vacation” etc..

Anomaly Detection on a Bridge Produced by Trains of Very High Speed

Anomaly detection on a bridge from TGV (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

Construction du Tableau de données symboliques

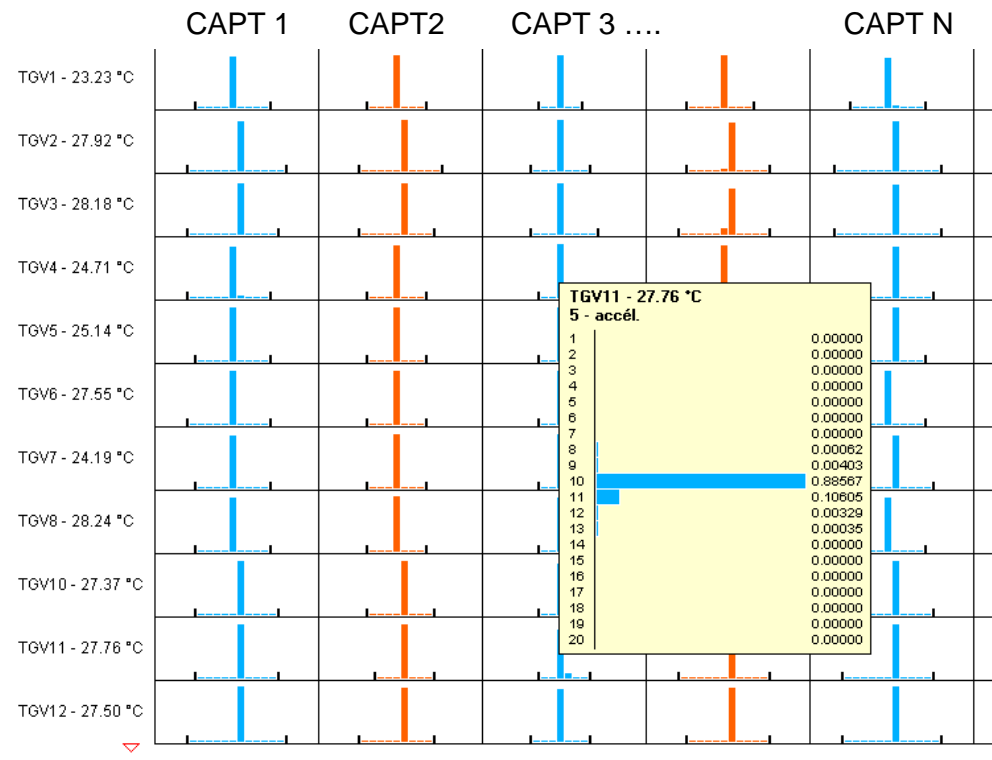
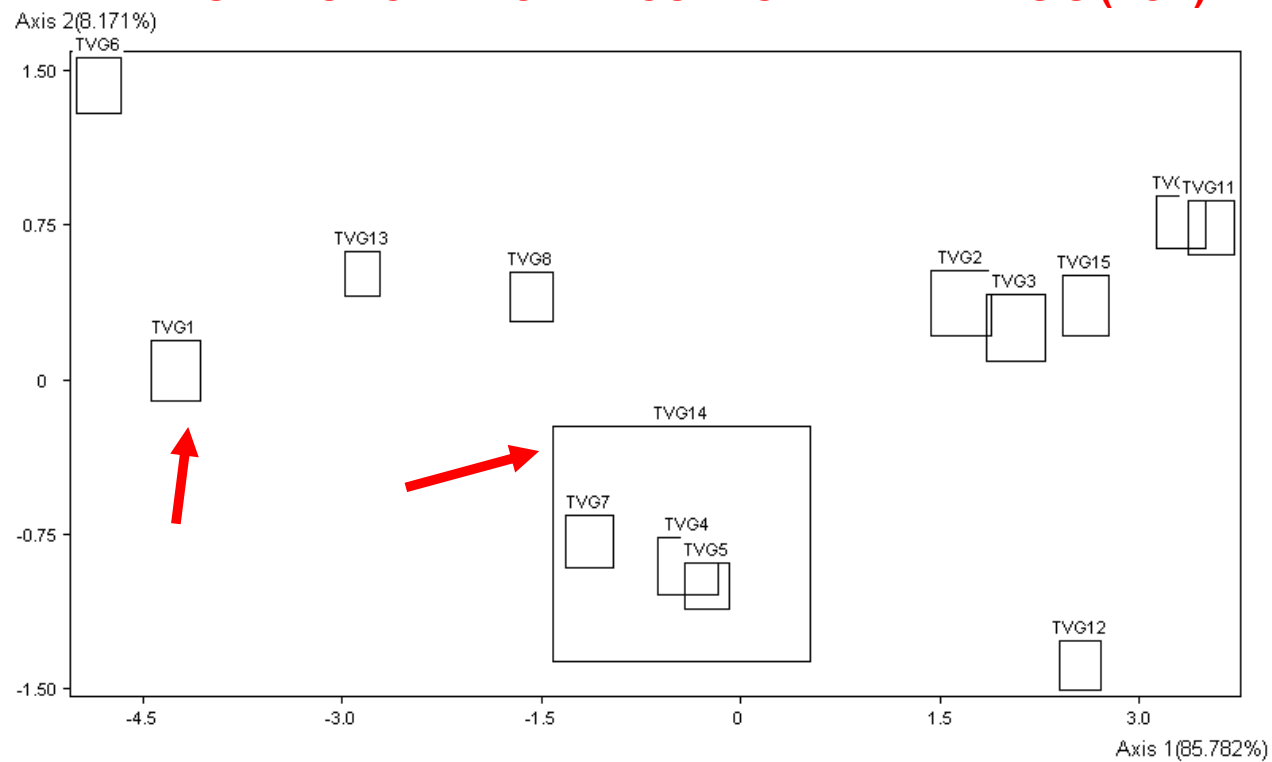


Tableau de données symboliques dont chaque case contient un histogramme à 20 intervalles représentant chaque signal pour chaque capteur.

SYMBOLIC PRINCIPAL COMPONENT ANALYSIS (PCA)

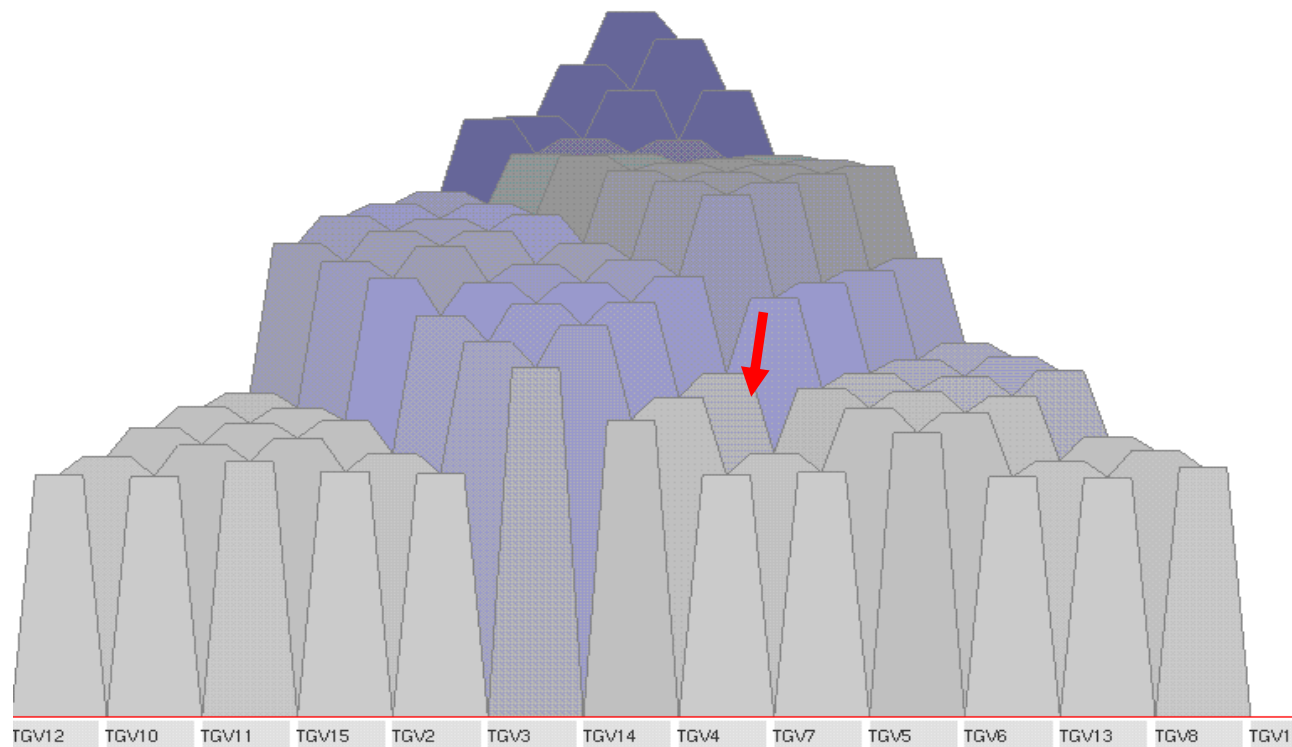


PCA on the interquartile intervals of the histograms contained in each cell.

Two anomalies are easily detected:

TGV1 is out of its group of temperature,

TVG14 covers all the trains of its group of temperature .



The symbolic pyramidal clustering confirms the anomalies.

- 1) TGV1 is out of its group of température
- 2) TGV 14 covers all the TGV of its group of température

Open directions of research

Practice

- Application to huge or very large data by symbolic preprocessing taking care of categories.
- Text mining in order to extract themes describing classes of documents by Symbolic Data

Open directions of research

Theory

- Galois Lattices are the underlying structure.
- Statistics: parametric or non parametric of SDA (ex : density of interval data, or histogram data,..)
- Partioning, hierarchies, overlapping clustering (Pyramids), Conceptual or based on dissimilarities between SD.

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 analysis to symbolic analysis.

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.