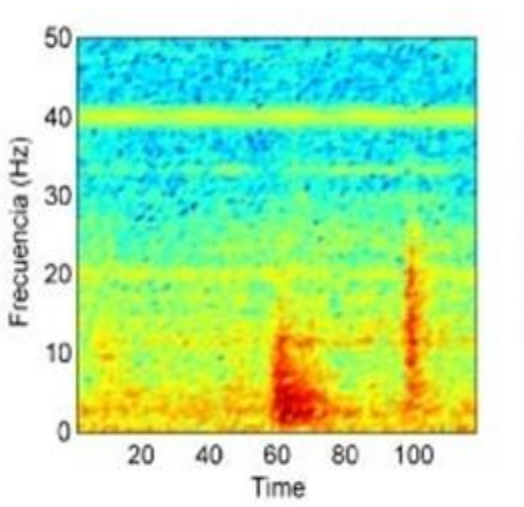
The Dissimilarity Representation as a Tool for Three-way Data Classification: a 2D Measure

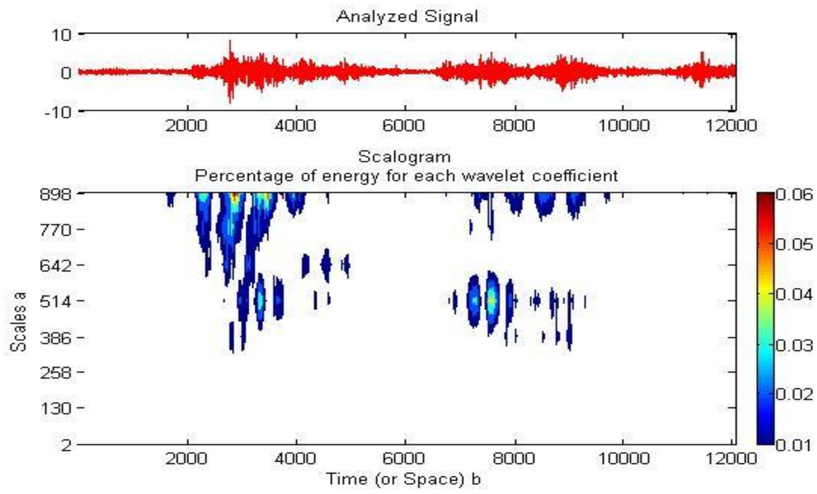
Authors: Diana Porro Muñoz
Robert P.W. Duin
Mauricio Orozco Alzate
Isneri Talavera Bustamante
John Makario Londoño Bonilla

S+SSPR 2010, Turkey

August 2010

Examples of multi-dimensional data

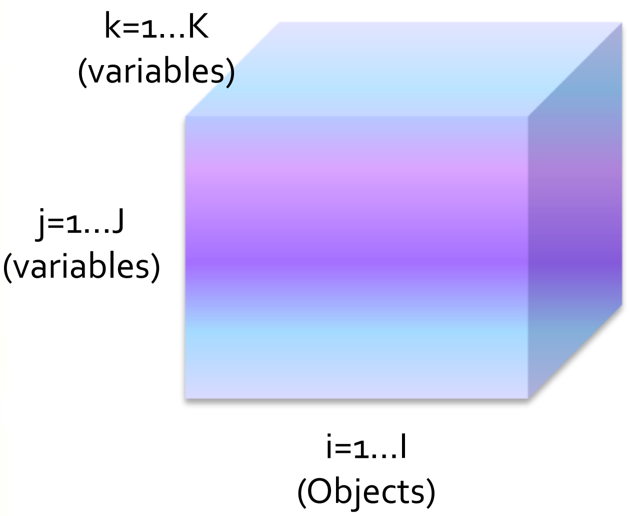


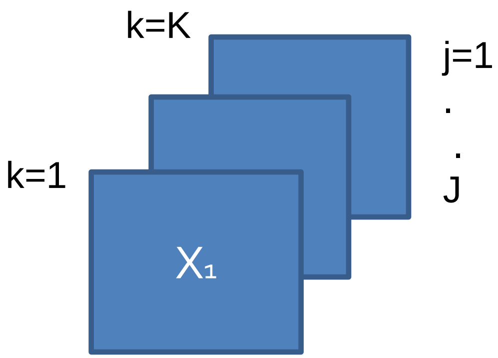


Multi-way data analysis

- So, an intuitive way to represent this relationship for all the objects would be in a three-way array

$$Y \in \mathbb{R}^{I\times J\times K}$$





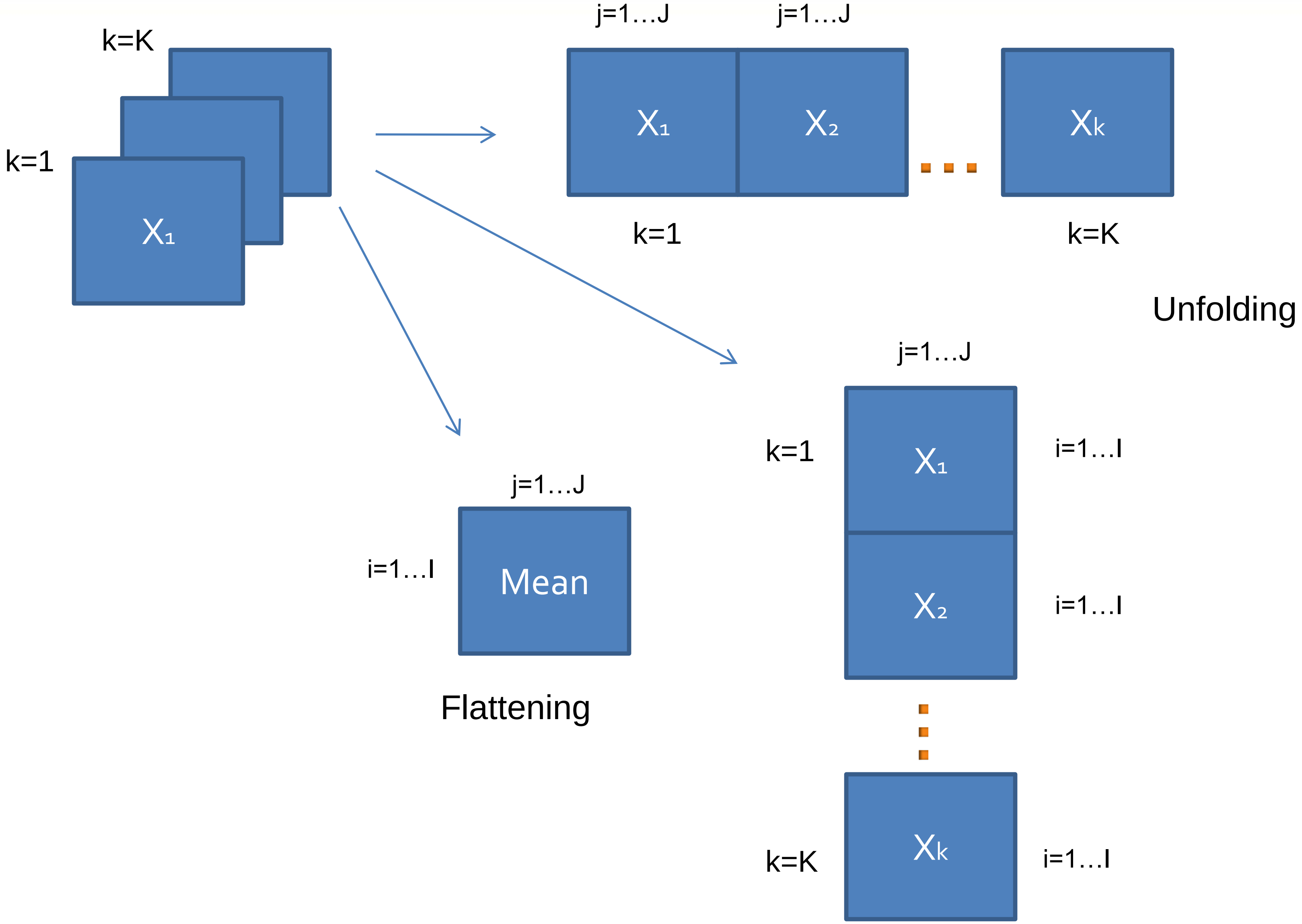
<u>Existing methods</u>

- Regression
- Exploratory analysis

Supervised classification????

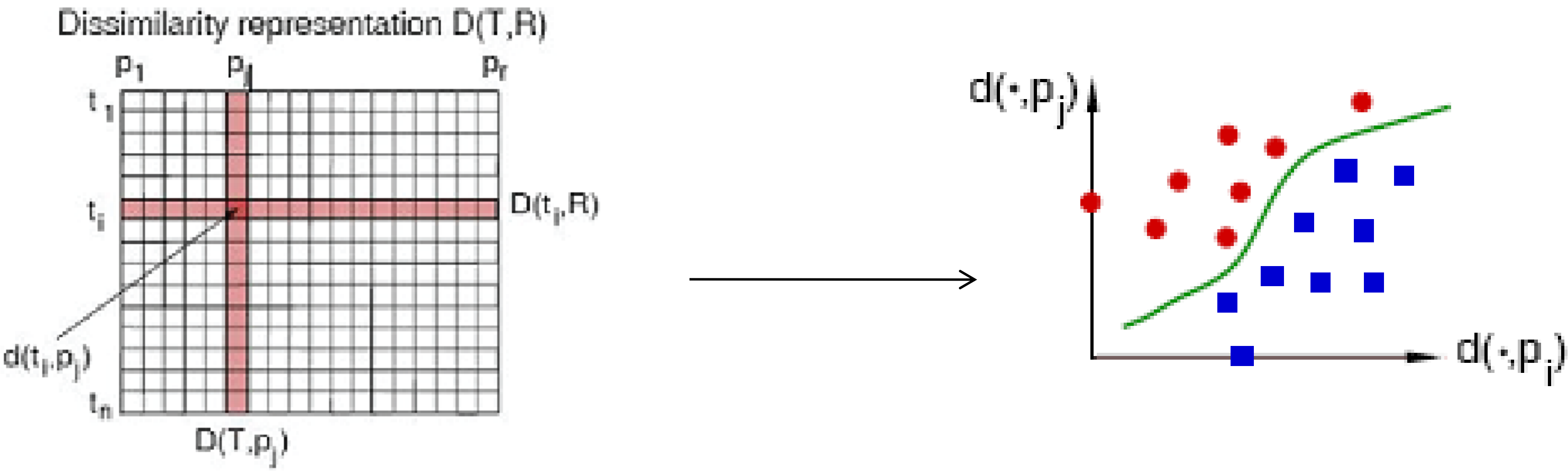
- Unfolding (1D) + classifiers

Unfolding


k=K
k=1
X1
j=1…J
j=1…J
X1
X2
Xk
k=1
k=K
Unfolding
j=1…J
k=1
X1
i=1…I
X2
i=1…I
k=K
Xk
i=1…I
j=1…J
i=1…I
Mean
Flattening

Dissimilarity Representation (DR)

It proposes to train classifiers in the space of the proximities between objects, instead of the traditional feature.



Now, the objects are represented by a dissimilarity matrix $D(T,R)$.

Objects are represented in this space by the column vectors of the dissimilarity matrix.

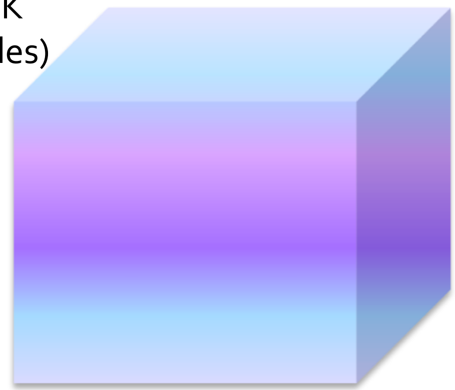
Some advantages of DR for our problem

- ✓ High dimensionality spaces
- ✓ Any knowledge or information about the problem background can be included into the dissimilarity measure.
- ✓ Any traditional classifier can work on the Dissimilarity Space!!!

+

DR can be generated from many initial representations of the objects eg. numerical vectors, graphs, as long as a "suitable" measure is used.

DR to classify three-way data?



$$Y \in \mathbb{R}^{I\times J\times K}$$

$$y_i \in \mathbb{R}^{J\times K}$$

?

h=1…r
Prototypes (Representative objects)

i=1…I
(Objects)

$$\begin{bmatrix} d_{11} & d_{12} & d_{13} & \cdots & d_{1r} \\ d_{21} & d_{22} & d_{23} & \cdots & d_{2r} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{l1} & d_{l2} & d_{l3} & \cdots & d_{lr} \end{bmatrix}$$

DR for multi-way data

For a t-dimensional array $Y \in \mathbb{R}^{I_1 \times I_2 \times \ldots, \times I_t}$, the theory of the DR holds.

Originally, each object is represented by a (t-1)-dimensional array, and all the objects together compose the t-dimensional array.

Consequently, to obtain the dissimilarity space, it is just done a mapping $\phi(\cdot, R) : \mathbb{R}^{I_1 \times I_2 \times I_{t-1}} \to \mathbb{R}^r$, such that for each object it is obtained

$$\phi(x_i, R) = \left[d(x_i, p_1), d(x_i, p_2), \ldots, d(x_i, p_r) \right]$$

2D Measures for three-way data

AMD : $$D^1_{a,b} = \left(\sum_{k=1}^{K} \left(\sum_{j=1}^{J} \left(y_{a,j,k} - y_{b,j,k} \right)^2 \right)^{p/2} \right)^{1/p}$$

- p<1 all the differences will be reduced. Larger ones will not interfere much in the measure.
- p>1, the larger differences will be more pronounced, resulting in a heavy influence on the measure.

2D Shape measure for three-way spectral data

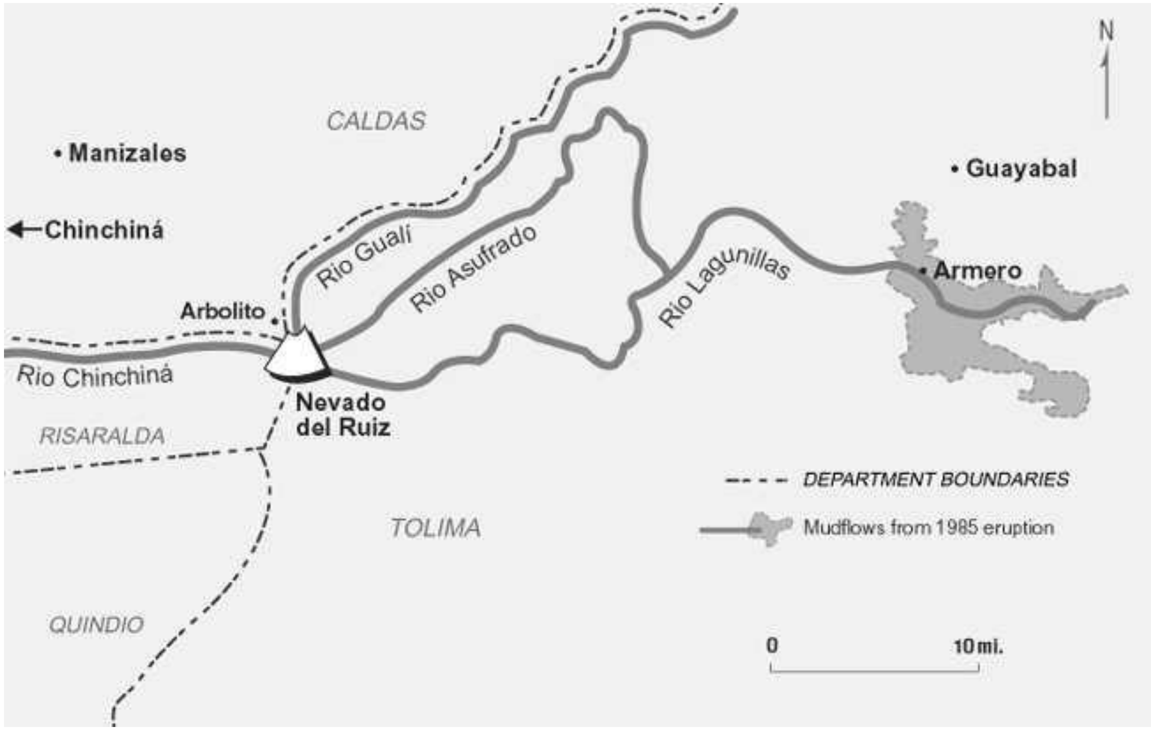
$$D_{a,b}^{1} = \left(\sum_{k=1}^{K} \left(\sum_{j=1}^{J} \left(\overset{\sigma}{y}_{a,j,k} - \overset{\sigma}{y}_{b,j,k} \right)^{2} \right)^{p/2} \right)^{1/p}, \quad y_{i,j,\cdot}^{\sigma} = \frac{d}{d_j} G(j,\sigma) * y_{i,j,\cdot}$$

$$D_{a,b}^{2} = \left(\sum_{j=1}^{J} \left(\sum_{k=1}^{K} \left(\overset{\sigma}{y}_{a,j,k} - \overset{\sigma}{y}_{b,j,k} \right)^{2} \right)^{p/2} \right)^{1/p}, \quad y_{i,\cdot,k}^{\sigma} = \frac{d}{d_k} G(k,\sigma) * y_{i,\cdot,k}$$

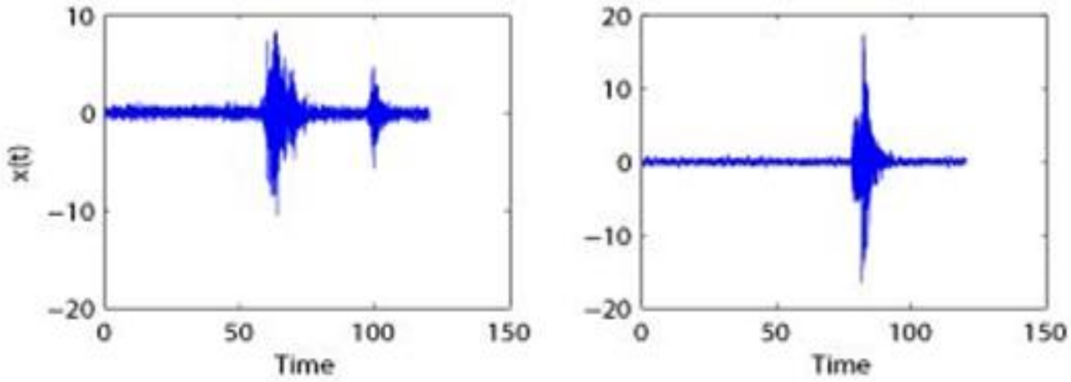
$$D = \frac{1}{w_1} D^1 + \frac{1}{w_2} D^2$$

Classification of seismic volcanic data

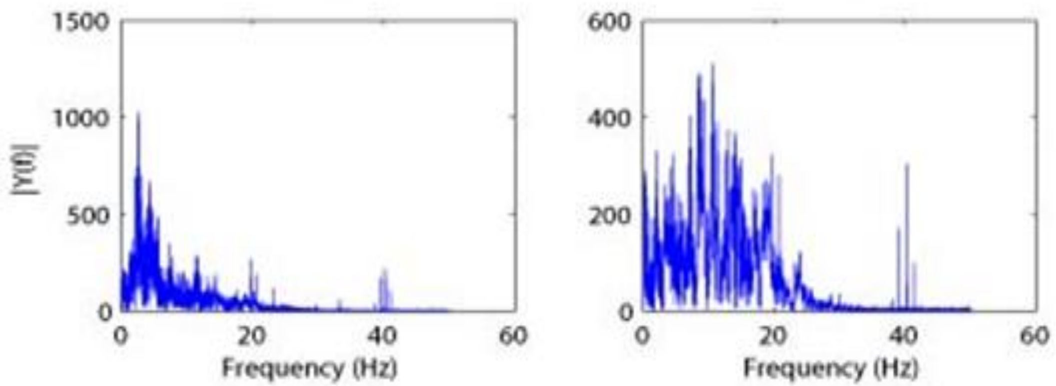
The automatic classification of seismic volcanic signals is an essential task nowadays, with the goal of discovering the interaction between volcanic earthquakes and volcanic processes.



Traditional Representation of seismic volcanic data



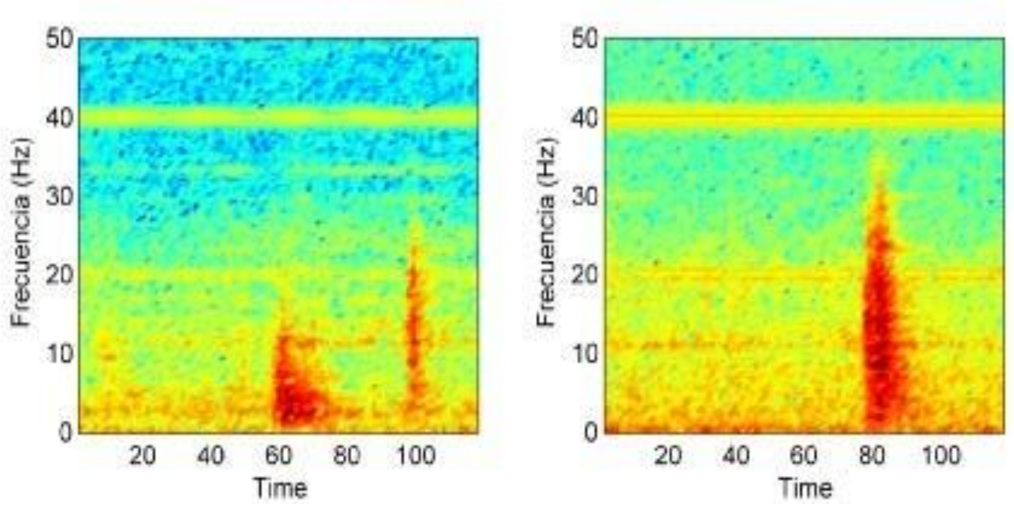
Does not take frequency changes along time into account



Does not take into consideration the time information

Time or frequency representations alone may not be optimal for seismic signal analysis, since spectral energy changes in time

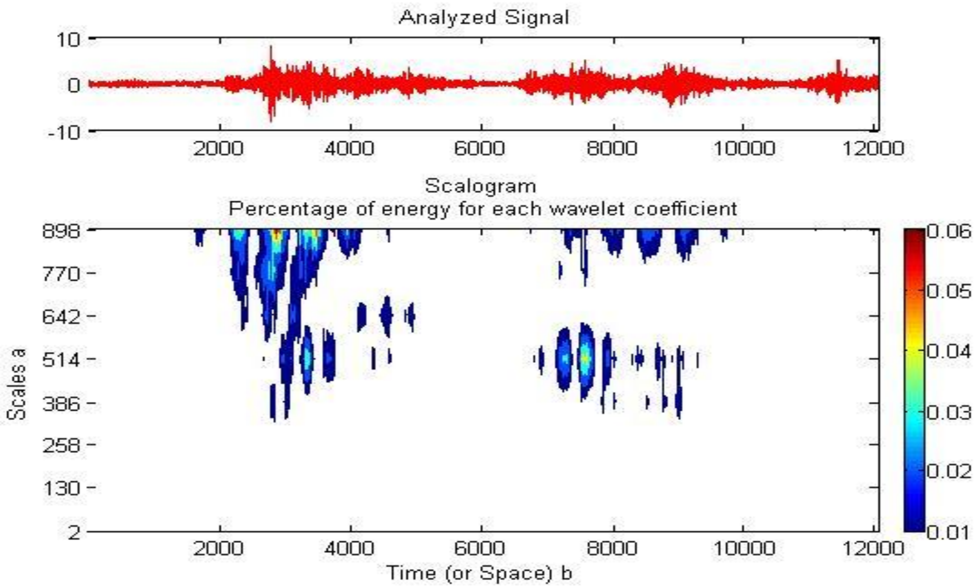
2D Representation. Spectrograms



- Spectrograms: Time-frequency representation , showing frequency changes in time.

- It can be known what frequency intervals are present in a time interval of the signal, but not with much precision.

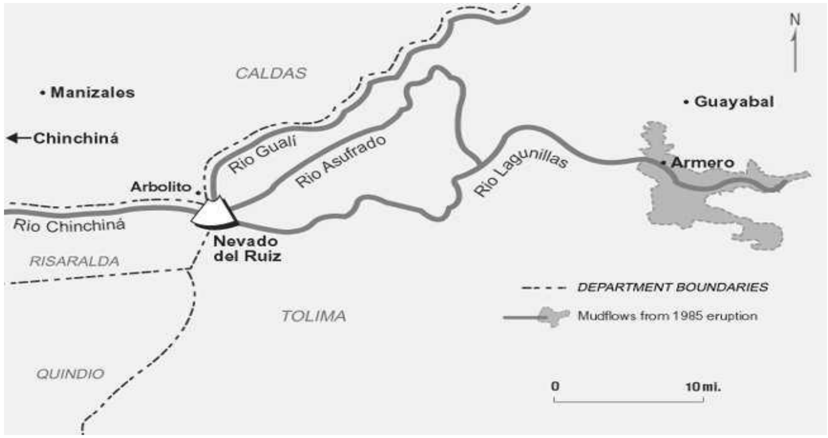
2D Representation. Scalograms by Continuous Wavelet Transforms (CWT)



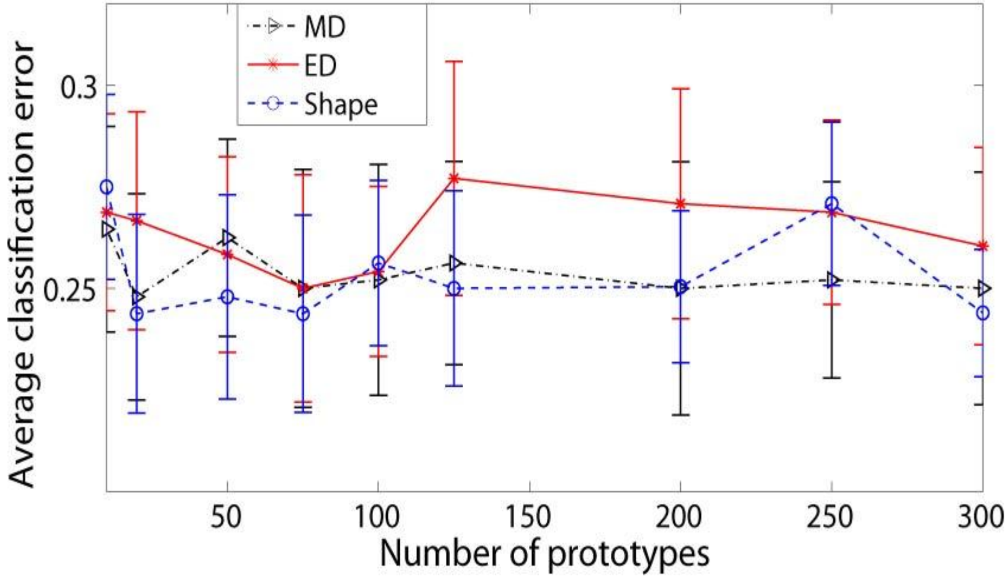
- The observations in the signal can be seen as continuous single entities, instead of sets of different variables.

Experiments with signals from Nevado del Ruiz Volcano

- Volcano: Nevado del Ruiz , Colombia
- Classes: Long Period, 235 events
  Volcano Tectonic earthquakes, 235 events
- Event: Time signals of 12032 points (120 sec.)
- 1D (spectral):12032-point Fast Fourier Transform (FFT).
- 2D (spectrogram): 256 short time Fourier transform
  Windows size: 256 points with 50\% of overlap.
- Data set: 470 x 129 x 93
- 2D (scalogram) : Morlet wavelet was used .
  Scale values: Major frequency components in the signals.
- Data set: 470 x 72 x 12032

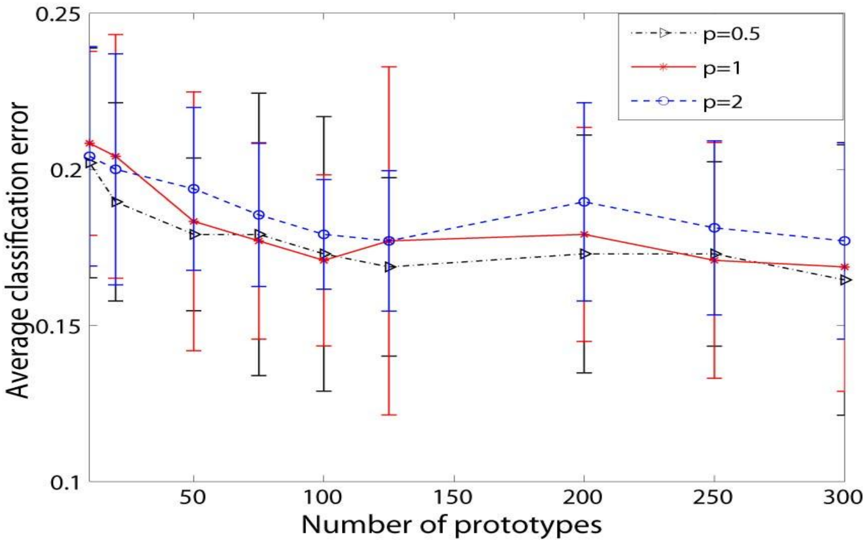


Results (1D representation)



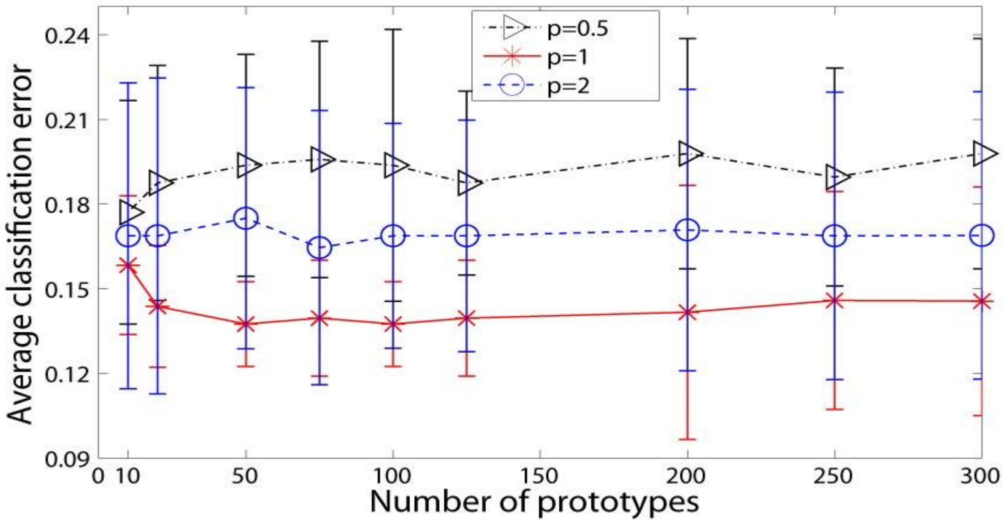
Measures: **1- Manhattan (MD)**
2- Euclidean (ED)
3- Shape

Results . Spectrograms (left) and Scalograms (right)



2DShape

AMD



Conclusions

- It was applied the Dissimilarity Representation as a tool for classifying three-way data.
- The relationship between the different dimensions is analyzed e.g. change of frequency content in time.
- The experiments results ratify our hypothesis that the 2D representation can be more discriminative if its structure is taken into account. Besides, the proposed 2D measure is capable of capturing this information.

Conclusions

- It was also shown the evidence the importance of the selection of a suitable dissimilarity measure for the problem at hand.

- Although this paper was more focused on the solution for three-way data, it can be extended to multi-way. Further studies will be done on this aspect.

Thank you!!!!!