ACI « Masse de Données » - 2003

DataHighDim

Analyse exploratoire et discriminante de données en grande dimension

Anne Guérin-Dugué

Laboratoire CLIPS – Grenoble

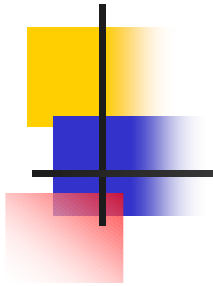
UJF, CNRS – UMR 5524

CLIPS

Communication Langagière et Interaction Personne-Système

Fédération IMAG

BP 53 - 38041 Grenoble Cedex 9 - France



Partenaires

- Lab. CLIPS-UJF (Coordinateur), Equipe MRIM, Grenoble
- Lab. LIS-INPG, Equipe SIC, Grenoble
- Equipe SELECT-INRIA FUTUR, Paris Orsay

- Lab. DICE-UCL, Equipe Machine Learning, Louvain-la-neuve, Belgique
- Lab LDG, CEA, Bruyères-le-Chatel

Objectifs

- Analyse Exploratoire et Discriminante de données en grande dimension
  - Méthodes et Outils
- Masse de données, Grande dimension ?
  - Plusieurs dizaines de milliers d'observations
  - Une centaine de variables (dimension)
- Données ?
  - Tableaux « observations x individus »
  - Tableaux croisées de dissimilitudes
  - Données manquantes

Adaptation des cartes de Kohonen aux données de dissimilitudes

Anne Guérin-Dugué
Sébastien Riou
Collaboration Corentin Dautricourt (UCL)

CLIPS
Equipe MRIM
BP 53 38 041 Grenoble Cedex 09

CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

Plan

- Position du problème
- Différentes approches
  - Positionnement Multidimensionnel
  - Extension aux dissimilitudes
- Algorithmes
- Résultats préliminaires
  - Données artificielles
  - Données réelles …
- Conclusion, Perspectives

Positionnement du problème

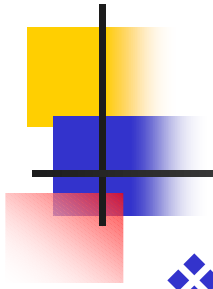
- Ensemble d'objets : $\aleph = \{o_i, i=1 \ldots N\}$
- Fonction de similarité ou bien de dissimilitude :
  - $S : \aleph \times \aleph \rightarrow \Re^+$

    $(O_i, O_j) \rightarrow D(O_i, O_j)$

- O_i n'est pas décrit par un ensemble de caractéristiques

- Analyse exploratoire non supervisée

Exemples d'application



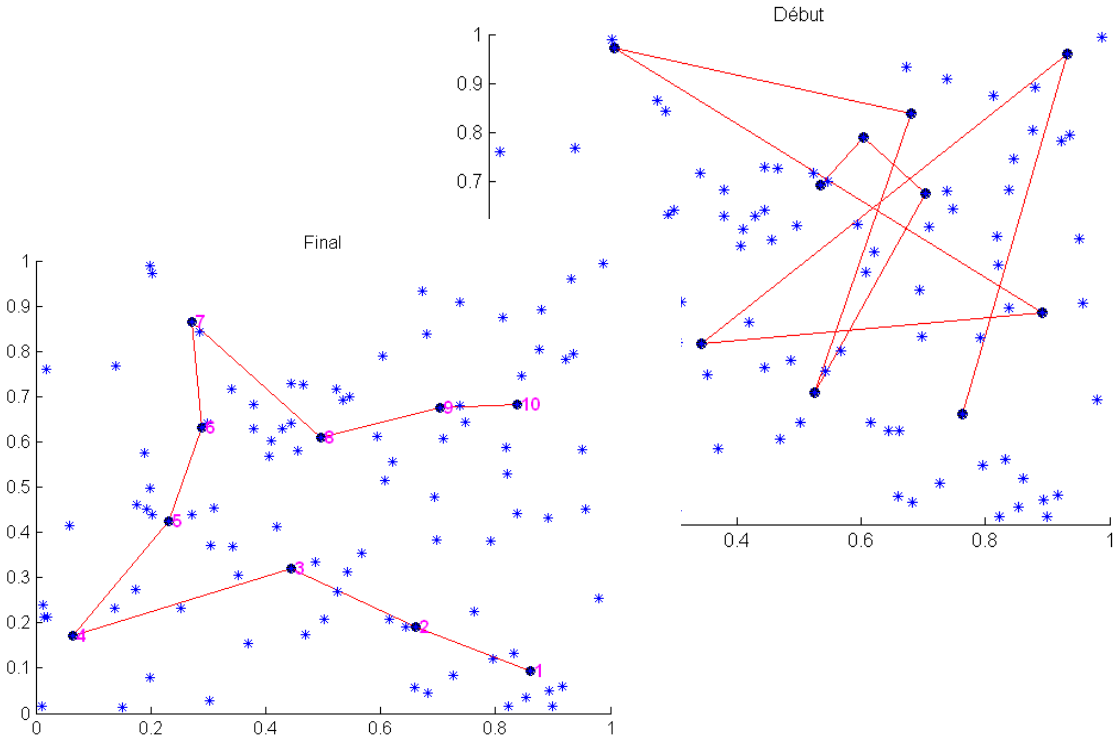
- Biologie, Génétique, …
- Psychophysique, …
- Finances (profil)
- Informatique : Usage du Web, …
- Traitement de l'information
- …

Cadre de l'étude : Analyse exploratoire par SOFM (Self-Organizing Feature Map)

X(n,p) : N points en p dimensions

→

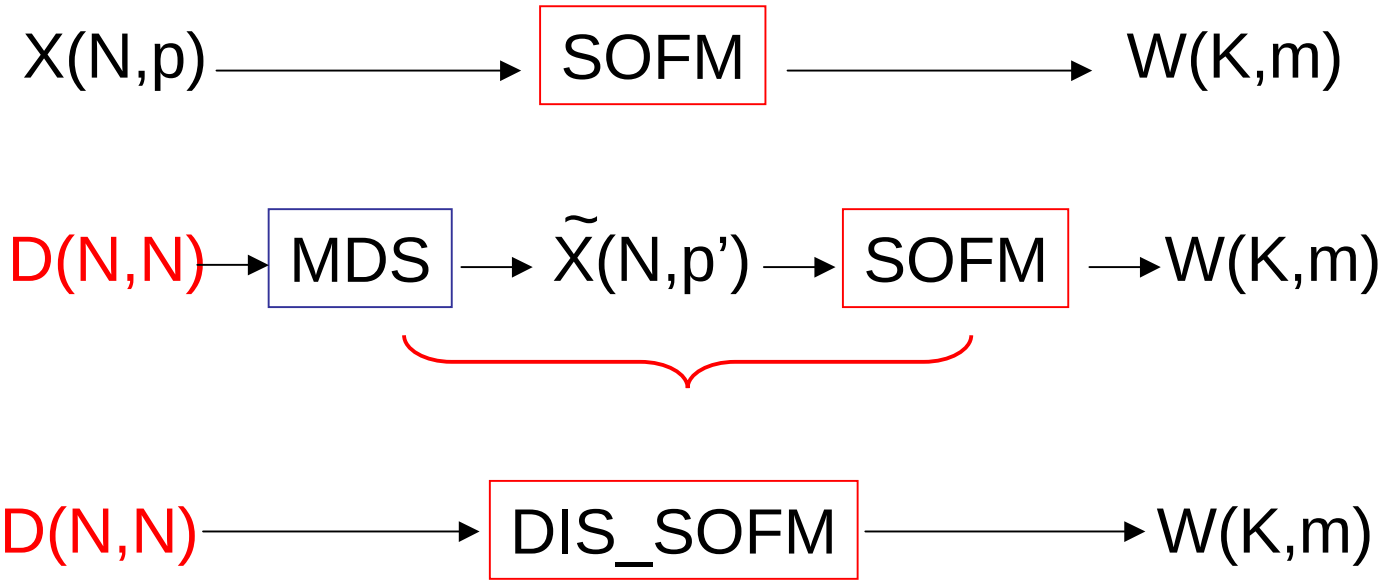
W(K,q) : K prototypes en q dimensions



Réduction d'information :
1/ nombre de dimensions (q <p)
2/ nombre de points (K < N)
en respectant au mieux la topologie d'entrée

Les alternatives

- Cas usuel : Représentation euclidienne
- Cas d'étude : Représentation par dissimilitudes



Bibliographie

- Bibliographie importante
  - Utilisation en génomique, texte (text mining), Web (web mining)
    - Proximité par « string edit distance », « Dynamic Time Warping », …
    - [Kohonen 1998 ; Kohonen, Somervuo, 2002 ; Somervuo, 2004]
  - Analyse de données symboliques
    - [El Golli, Conan-Guez, Rossi, 2004]
  - Approche issue de mélange de gaussiennes
    - [Ambroise, Govaert, 1996]
  - Approche stochastique, recuit simulé
    - [Seo, Obermayer, 2004]

Principe de l'extension aux dissimilitudes

- Algorithme de départ
  - SOFM en version « Batch »

Algorithme : SOFM « Batch »

- Initialisation (t=0)
  - Structure de la carte (K prototypes, q dim)
  - Initialisation des prototypes $\vec{\omega}_k$ en entrées :
- Affectation (t)
  - Partition de X par minimum de distance aux prototypes
- Topologie (t)
  - Fonction de voisinage, distance sur la carte :
  - Diminution du rayon r
- Représentation :
  - Nouveau prototype :

$$f(i) = \arg\left[\min_k \left\{d^2(\vec{x}_i, \vec{\omega}_k)\right\}\right]$$

$$h_{kl} = \exp - \left\{\frac{d_c^2(k,l)}{2.r^2(t)}\right\}$$

$$\vec{\omega}_k = \frac{\sum_{i=1}^{N} h_{ik}.\vec{x}_i}{\sum_{i=1}^{N} h_{ik}}$$

Algorithme : Extension aux dissimilitudes
Exemple : Median SOFM

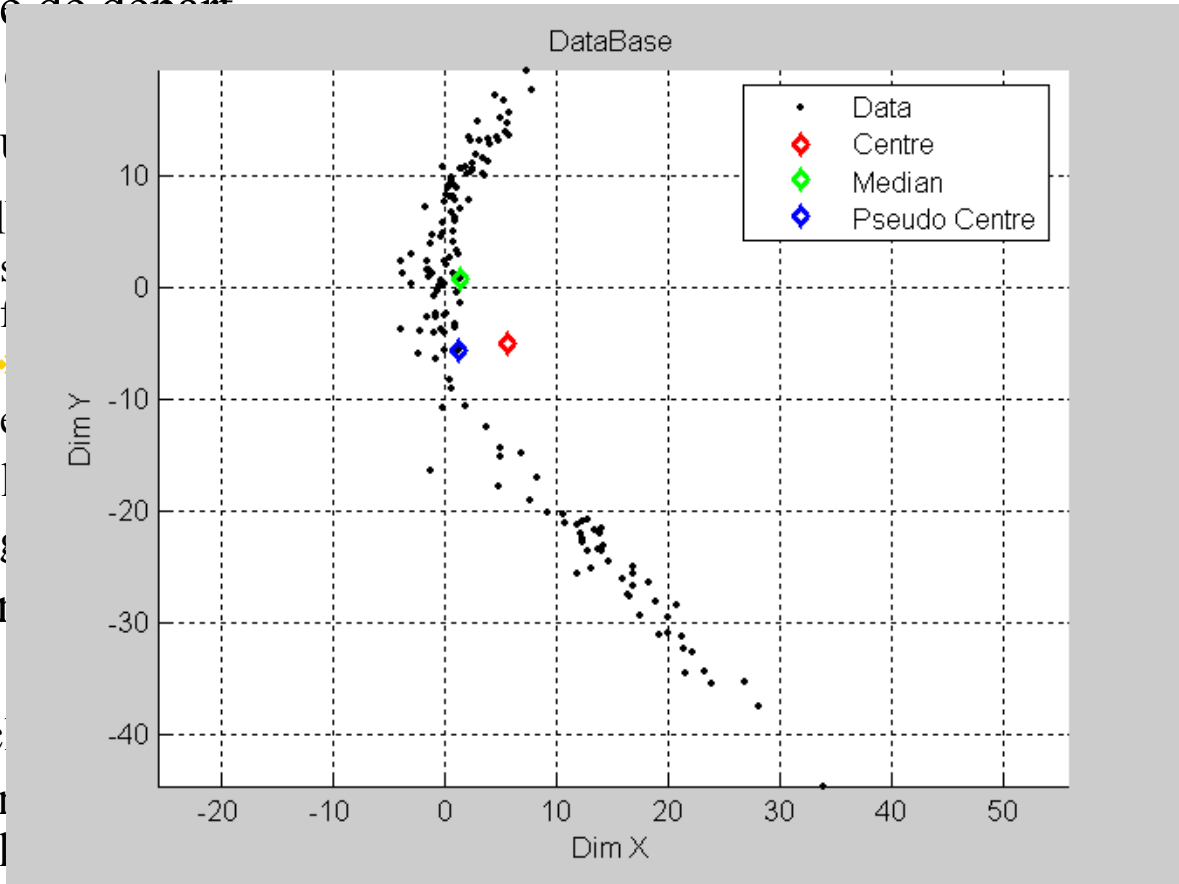
- [Somervuo, Kohonen] :
  - Affectation : A partir des dissimilitudes d'entrée
  - Topologie : Idem
  - Représentation :
    - calcul barycentrique direct → Impossible
    - Pour chaque prototype k, Recherche par optimisation l'observation « référente » (set médian) $\omega_k \in O$ (set médian) minimisant :

$$\sum_{o_i \in O} h_{kf(i)} d(\omega_k, o_i)$$

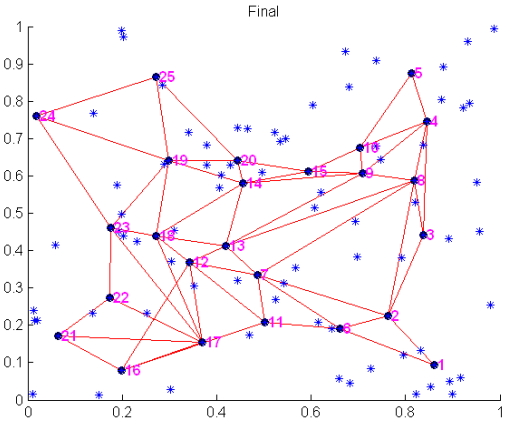
Principe de l'extension aux dissimilitudes

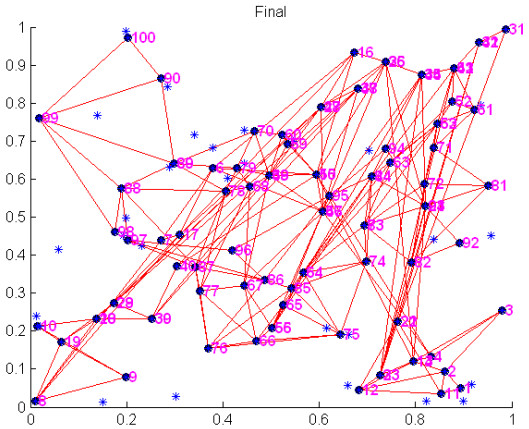
- Algorithme de départ
  - SOFM
- Mise en œu
  - Conséq
    - plus d'a
    - Erre
  - Difficul
    - Iné
- Propositio
  - Ne pas
  - Autre c
  - Intégre triangul

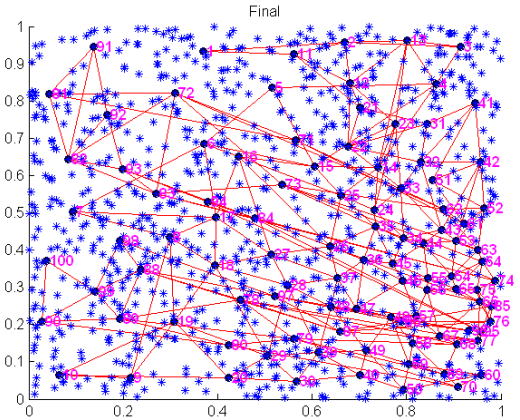
pe

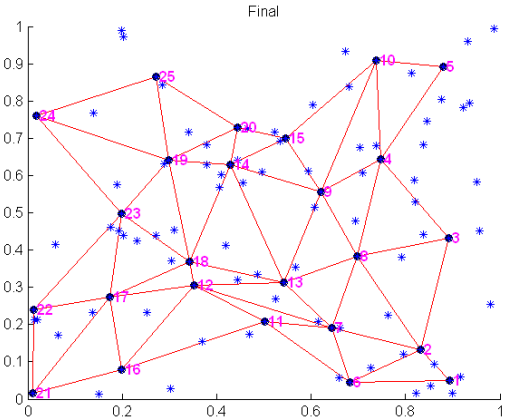


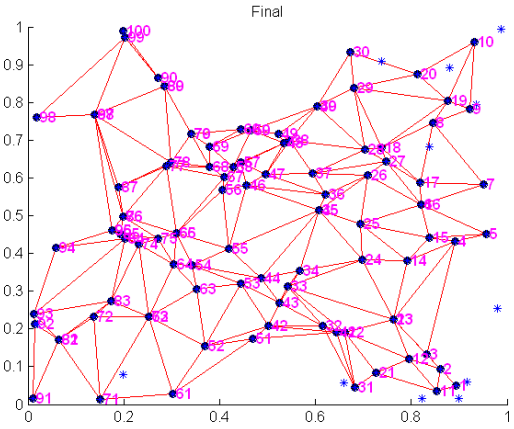
Données artificielles euclidiennes

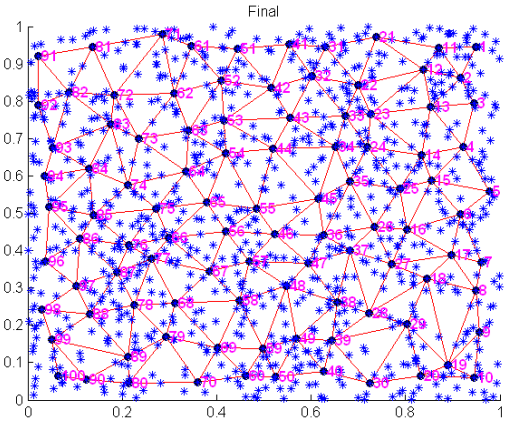






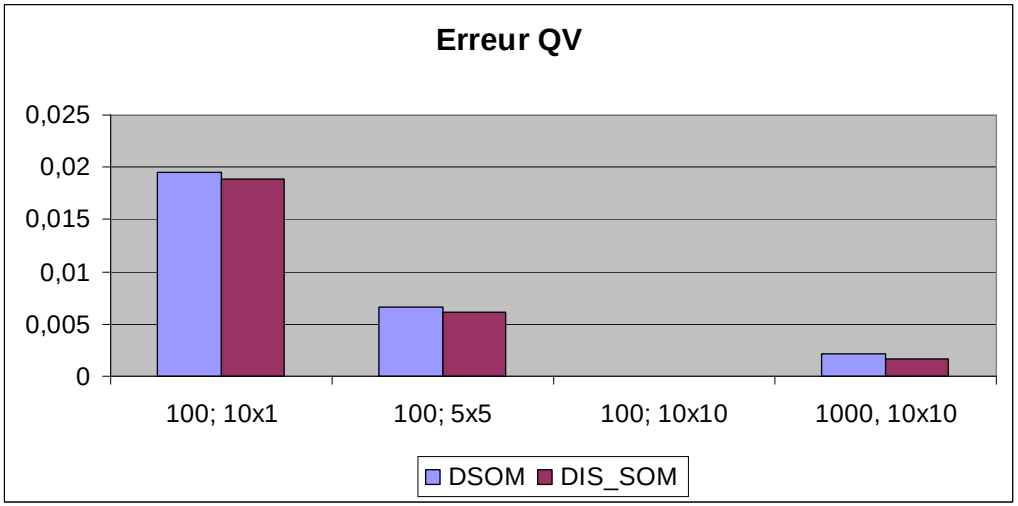


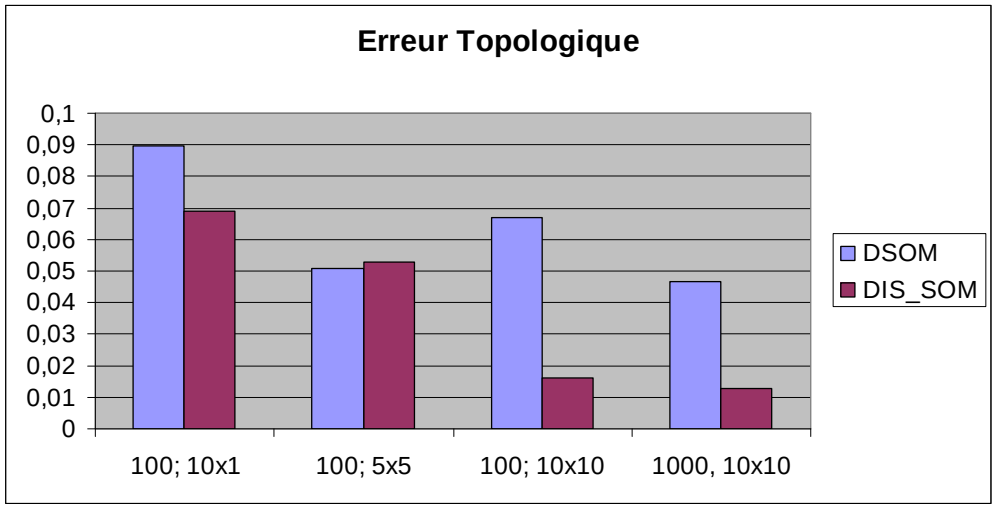




Données artificielles euclidiennes

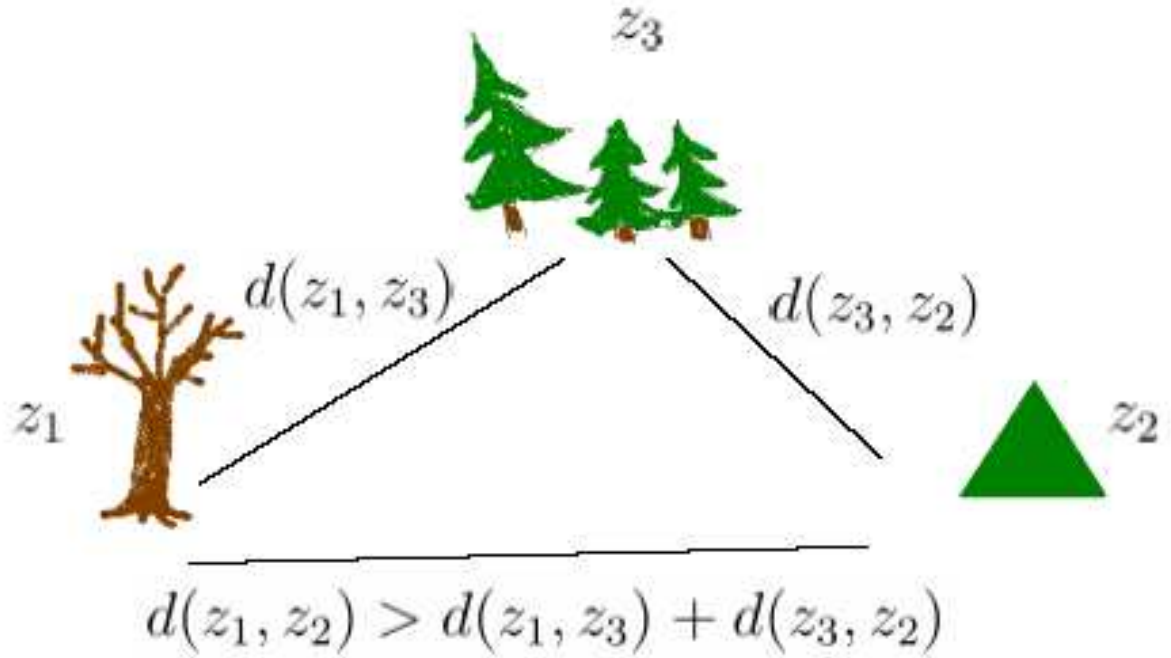
- Tirage aléatoire
- D= distance euclidienne





Données artificielles non euclidiennes

- Données non euclidiennes
  - Inégalité triangulaire non respectée
    - Savoir si elle est « fortement » ou « faiblement » non respectée
    - On peut observer une croissance de l'erreur d'une itération à l'autre
    - Pénalisation dans l'étape de représentation
  -



Résultat DSOM

c3_4_1 c4 c3 c3_4 c3 c3 c3 c3 c1_3 c1_3 c3 c1 c3 c4 c3	c1 c1 c1 c1	c4_2_1 c1 c4 c1 c4	c1	c2 c2 c2 c2
c1	c4_2_1 c4 c1 c1_3_2 c1 c4 c4 c1 c1 c1 c4_2 c4 c1 c4	c1	c2 c2 c2	c2
c2_1_4 c4 c1 c1 c4_1 c4_2 c1_2 c4 c1_4 c4_1	c1	c1	c2	c2
c2_1_4 c4 c4_1 c1 c4 c4 c4_1_2 c4	c1	c1 c1 c1	c1	c2_3_4 c2 c4 c4_2_3 c3 c4 c2_3 c4_3 c3_4 c3 c3 c2_3 c4_3_2 c3_4 c3 c3 c4_3 c3_2 c2_3 c3 c3 c3_4 c4
c2	c1_2 c1 c1 c1	c1 c1 c1	c1	c1

Résultat Dis_SOM

c1 c1 c1 c1 c1 c1_2 c1_2	c1 c1 c1 c1	c1 c1 c1 c1 c1	c2 c2 c2	c2 c2 c2 c2
c1 c1 c1	c1 c1 c1 c1 c1	c1 c1 c1	c2 c2	c2 c2 c2
c4 c4	c4 c4 c4_3 c4_3 c3_4 c3_4 c3_4 c3_4 c4_3_2 c2_3_4 c4_2_3 c3_4_1	c4 c4	c3 c3 c3	c3 c3 c3
c4 c4 c4_1 c4_1 c1_4 c4_1	c4	c4 c4 c4	c3 c3 c3	c3 c3 c3_2 c2_3 c2_3 c2_3 c1_3_2
c4 c4	c4 c4 c4	c4 c4_2 c4_2 c4_2_1 c2_1_4 c4_1_2 c2_1_4 c4_2_1	c3 c3	c3 c3 c1_3 c1_3

Base d'Images

[Dautricourt, 2005]