Cartes auto-organisées de Kohonen et Clustering

Marie Cottrell[1] & Nathalie Villa-Vialaneix[2]

[1] Université Paris 1 Panthéon-Sorbonne, laboratoire SAMM
[2] INRA Toulouse, unité MIAT

20 octobre 2015 - Journée Clustering

Données numériques vectorielles

Données dans un espace $\mathscr{X}$ de dimension p

- Espace muni d'une densité f, cas continu
- ou
- Espace fini contenant N données, cas discret

Les données peuvent être stockées, ou être disponibles on-line

On cherche à regrouper les données en classes, disjointes, bien séparées et homogènes

Algorithme de Forgy (CM)

Algorithme des centres mobiles ou **Algorithme de Forgy**
Le nombre de classes est fixé = K

- Algorithme itératif, **déterministe**
- Recherche de "centres"
- Initialisation aléatoire des "centres"
- Chaque étape est double
  - Définition des classes par les plus proches voisins
  - Mise à jour des "centres" : centres de gravité des classes

RQ de vocabulaire :
centres : représentants : vecteurs codes : prototypes

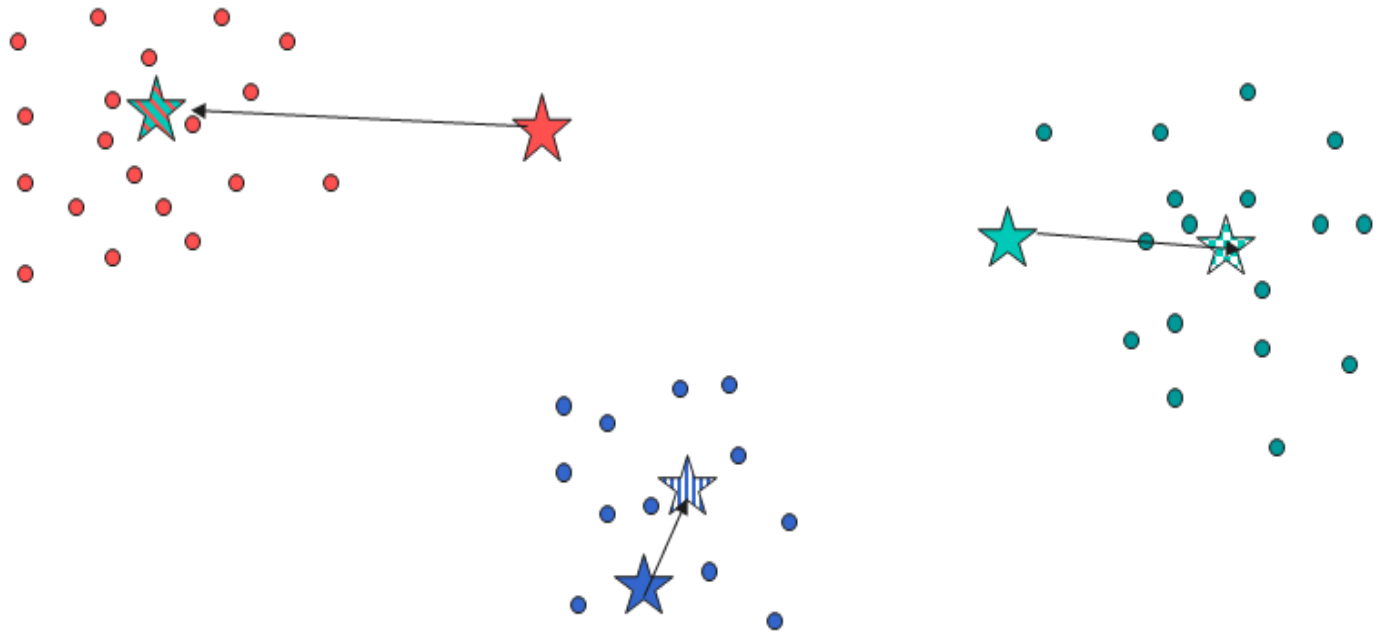


FIGURE : Une itération de l'algorithme de Forgy

Quantification Vectorielle (VQ ou SCL)

- Algorithme **stochastique**
- Initialisation aléatoire des prototypes
- Au temps t, si $m(t) = (m_1(t), \ldots, m_K(t))$ sont les prototypes, une donnée x est tirée au hasard
  1. Définition du numéro du prototype gagnant

$$c^t(x) = \arg \min_{k \in \{1,\ldots,K\}} \|x - m_k(t)\|^2, \quad (1)$$

  2. Mise à jour du prototype gagnant et de lui seul

$$m_c(t+1) = m_c(t) + \epsilon(t)(x - m_c(t)), \quad (2)$$

  où $\epsilon(t)$ est le gain d'apprentissage (positif, <1, constant ou décroissant)
- Algorithme adapté à l'apprentissage on-line

Les algorithmes de Forgy et VQ convergent vers un minimum local de la fonction **Distorsion**

$$E(m) = \frac{1}{2N} \sum_{i=1}^{N} \| m_{c(x_i)} - x_i \|^2.$$

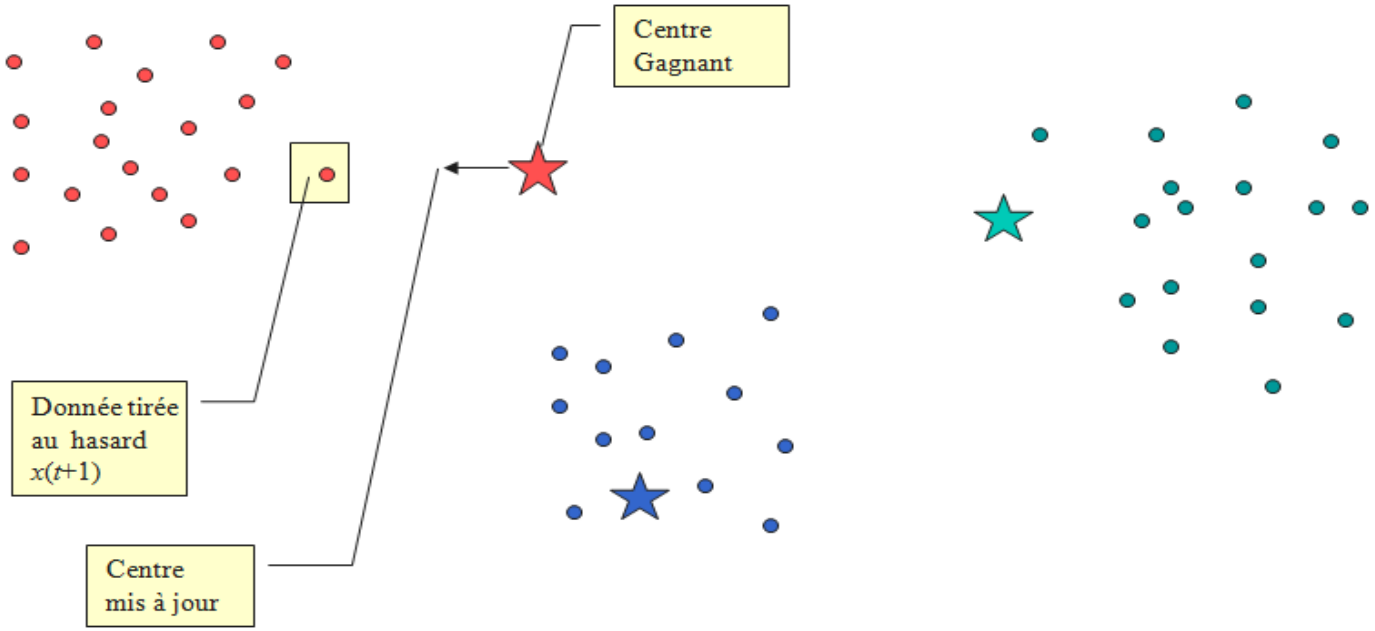


FIGURE : Une itération de l'algorithme de Quantification Vectorielle

Algorithme de Kohonen (SOM)

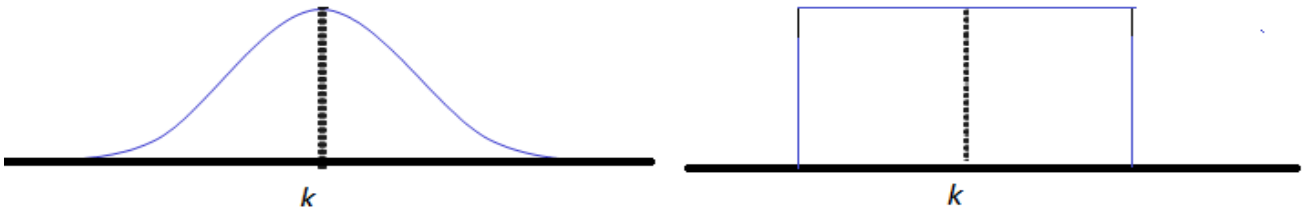
[Kohonen, 1982a, Kohonen, 1982b, Kohonen, 1995]

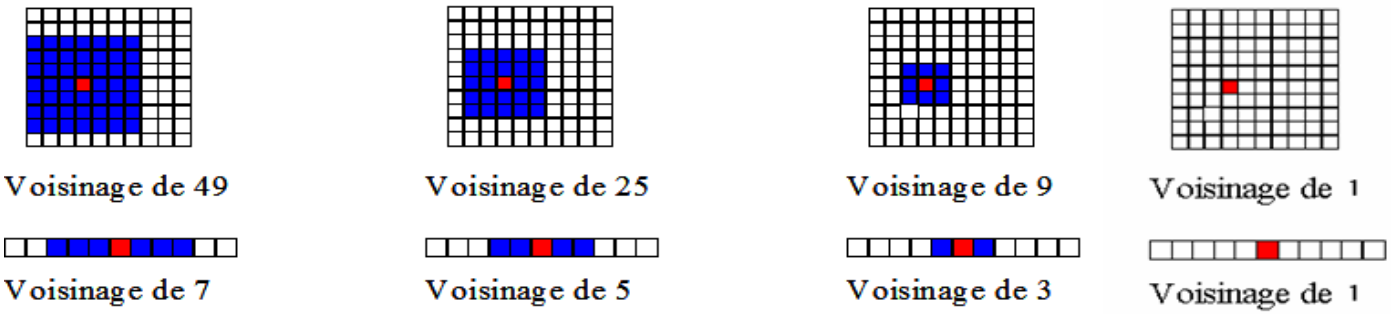
On introduit des voisinages entre les prototypes

Si $\mathcal{K} = \{1, \ldots, K\}$ est l'ensemble des indices des prototypes, disposés sur une grille ou une ficelle, on définit une fonction de voisinage sur les couples d'indices :

- $h_{kk} = 1$, h symétrique
- h_{kl} ne dépend que de la distance entre les indices k et l

Plusieurs choix possibles, les plus classiques :





Définition de l'algorithme de Kohonen (SOM)

K est le nombre de prototypes, la structure du réseau est une ficelle ou une grille, h est la fonction de voisinage

- Algorithme **stochastique**
- Initialisation aléatoire des prototypes
- Au temps t, si $m(t) = (m_1(t), \ldots, m_K(t))$ sont les prototypes, une donnée x est tirée au hasard
  1. **Définition du numéro du prototype gagnant**

$$c^t(x) = \arg \min_{k \in \{1,\ldots,K\}} \|x - m_k(t)\|^2, \tag{3}$$

  2. **Mise à jour du prototype gagnant et de ses voisins**

$$m_k(t+1) = m_k(t) + \epsilon(t) h_{kc^t(x)}(t)(x - m_k(t)), \tag{4}$$

où $\epsilon(t)$ est le gain d'apprentissage (positif, <1, constant ou décroissant)

Définition des classes

- Classe C_k : ensemble des données plus proches du prototype m_k que de tous les autres
- Partition (ou mosaïque de Voronoï) avec une structure de voisinage

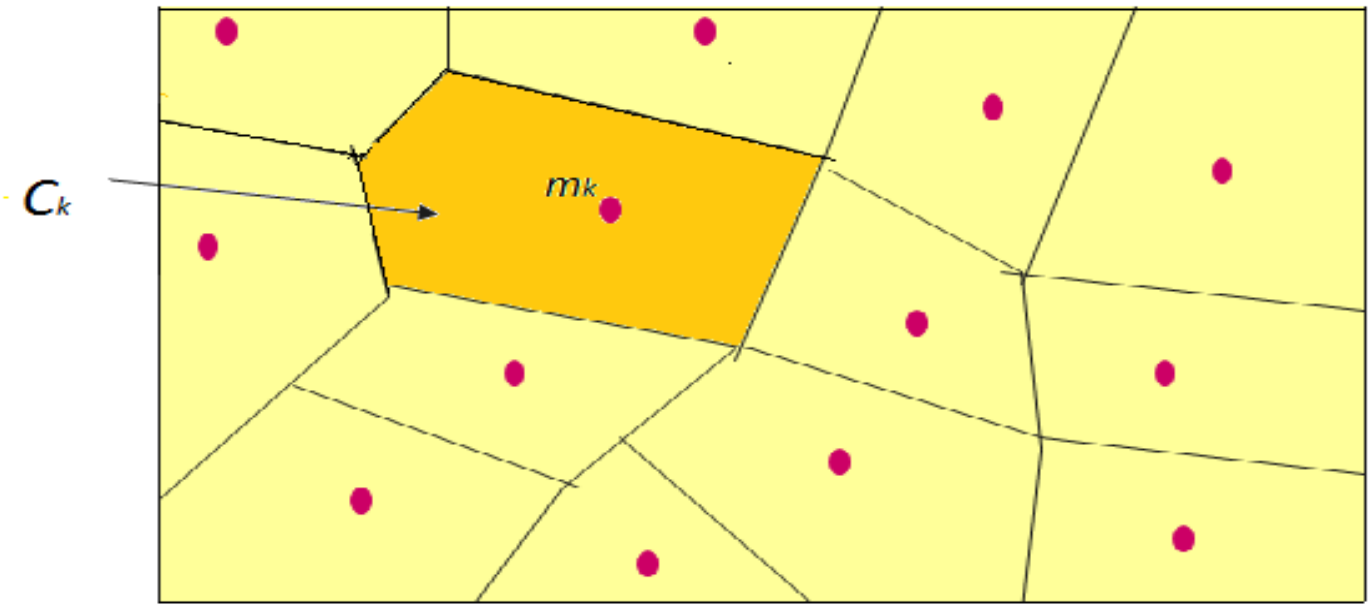


FIGURE : Mosaïque de Voronoï

La donnée $x \in C_k \iff m_k$ est le gagnant

SOM Batch

[Kohonen, 1995]

- Généralisation de l'algorithme de Forgy
- Algorithme **déterministe**
- Initialisation aléatoire des prototypes
- Chaque étape est double
  - Définition des classes par les plus proches voisins
  - **Mise à jour des prototypes** : centres de gravité d'un ensemble formé par la classe et les classes voisines (pondérées par la fonction de voisinage)

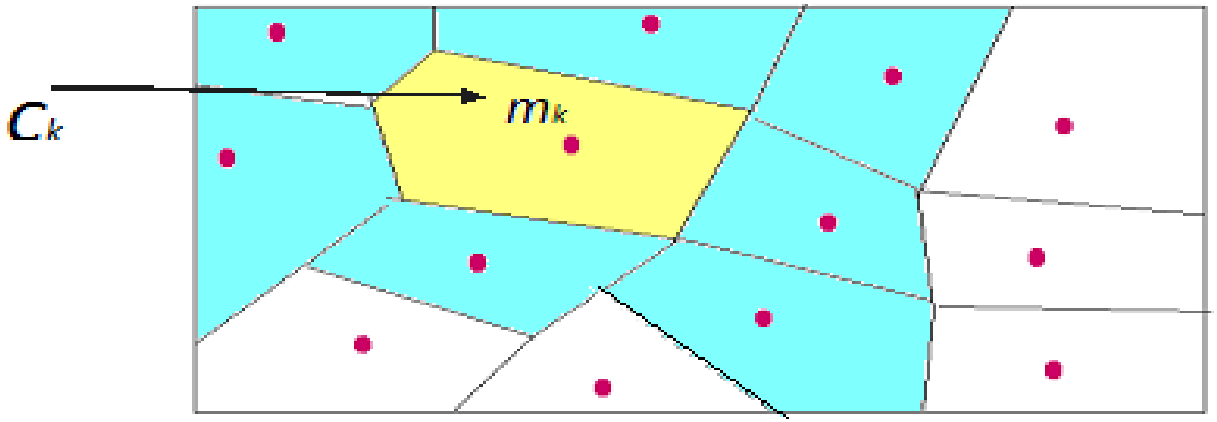


FIGURE : Algorithme Batch SOM

Relations entre les 4 algorithmes

	Stochastique	Déterministe
0 voisin	VQ, SCL	Forgy, centres mobiles
Avec voisins	**SOM**	SOM Batch

- SOM et SOM Batch **organisent : des données proches se trouvent dans la même classe ou dans des classes voisines**
- D'où les propriétés de **visualisation** des **cartes de Kohonen**, que n'ont pas CM et VQ
- SOM dépend peu de l'initialisation, alors que SOM Batch y est très sensible
- SOM Batch est **déterministe** et plus souvent utilisé dans les applications industrielles

35 pays en 1996, croissance de population, mortalité infantile, analphabétisme, scolarisation secondaire, chômage et inflation

Pays	Ancrx	Txmort	Txanal	Scol2	Pnbh	Chomag	Inflat
Afghanistan	6	159	70,9	15	276	19	17
Afrique du sud	2,6	46,9	23,5	77	2873	33	10
Albanie	1,1	33,1	8	29,2	828	17,9	16,1
Algerie	2,2	42,1	42	61	1570	27	31
Allemagne	0,2	5,8	1	101,2	24993	9,4	3,1
Angola	3,6	126,8	58	14	575	25	951
Arabie Saoudite	3	68,8	39,5	49	7081	6,6	0,7
Argentine	1,1	33,8	4,4	72,3	7827	11,3	4
Australie	1,3	5,9	0,1	84	17688	9,7	2,5
Bahrein	2,5	24,2	17	99	7500	15	2
Belgique	0,1	7,8	0,3	103,2	22225	12,6	2,6
Bolivie	2,2	74,9	20	37	733	6,2	8,5
Bresil	1,6	59,8	18	43	3073	5,5	1094
Bulgarie	-0,2	15,3	2,1	68,2	1058	17	33
Cameroun	2,9	85,8	36,5	32	733	25,1	12,8
Canada	1	6,7	3,1	104,2	18286	10,4	0,3
Chili	1,4	14,4	5,7	67	3643	6,1	11,2
Chine	1	25,8	22,4	55	418	2,5	22
Chypre	1	9,9	4,5	95	9459	2	4,8
Colombie	1,7	36,8	8,5	62	1379	8	22,9
Comores	3,5	81,7	42,5	19	317	16	24,8
Coree du Sud	1	14,9	3,7	96	7572	2,3	6
Costa Rica	2,2	13,5	5,2	47	1896	5	15
Cote d' Ivoire	3,3	90,9	46,8	25	587	17	25,6
Croatie	0,1	11,5	3,2	83,2	2755	13,1	97,6
Danemark	0,2	5,6	1	114,2	28346	12,1	2,1
Egypte	1,9	58,8	50,5	76	632	20,2	8,3
Emirats arabes unis	2,2	23,2	20,9	89	23809	0,2	5
Equateur	2,1	36,8	12,8	55	1205	7,2	26
Espagne	0,2	7,3	7,1	110,2	12283	24,4	4,8
Etats Unis	1	8,2	3	97,2	25219	5,6	2,8
Fidji	1,6	26,9	9	64	2077	5,5	1,5
Finlande	0,3	5,8	0,1	119,2	18803	18,4	2,2
France	0,4	6,3	1,1	106,2	22700	12,5	1,7
Ghana	3	82,9	58,8	36	404	11	25
Grece	0,7	8,4	4,6	99,3	7390	9,8	11
Guyana	1,1	48,8	1,9	17	243	13	15,4
Haiti	2,1	108,8	70	22	241	50	50
Hongrie	-0,4	13,9	1	81,3	3411	10,5	19,2

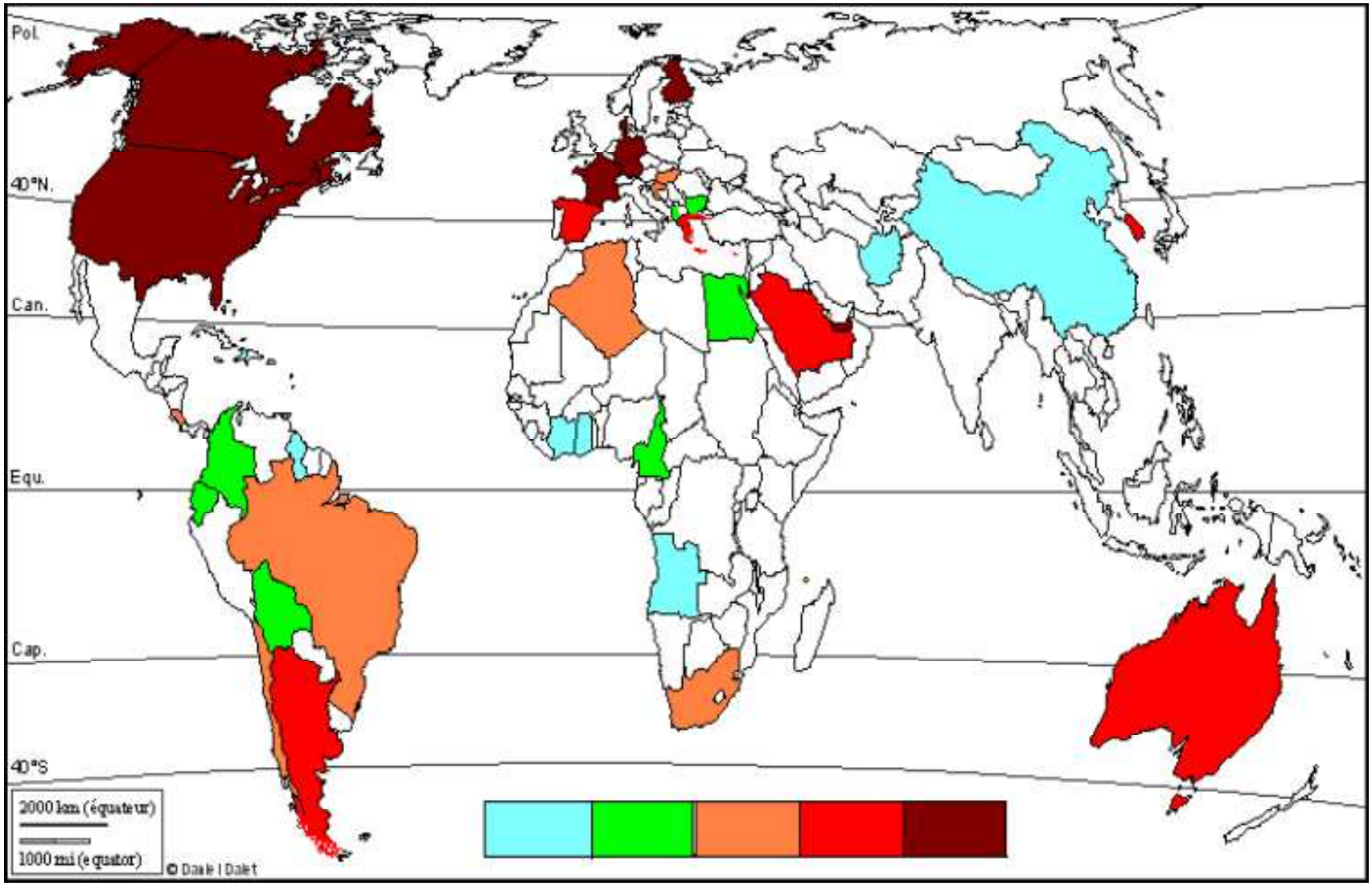
TRIS SELON LES 7 DESCRIPTEURS

	Croissance de population	Mortalité infantile	Analph	Scolarisation	Tx chomage	Inflation
PNB	ANCRX	TXMORT	TXANAL	SCOL2	CHOMAG	INFLAT
Haiti	Afghanistan	Afghanistan	Afghanistan	Angola	Haiti	Bresil
Guyana	Angola	Angola	Haiti	Afghanistan	Afrique du sud	Angola
Afghanistan	Cote d' Ivoire	Haiti	Ghana	Guyana	Algerie	Croatie
Ghana	Ghana	Cote d' Ivoire	Angola	Haiti	Cameroun	Haiti
Chine	Arabie Saoudite	Cameroun	Egypte	Cote d' Ivoire	Angola	Bulgarie
Angola	Cameroun	Ghana	Cote d' Ivoire	Albanie	Espagne	Algerie
Cote d' Ivoire	Afrique du sud	Bolivie	Algerie	Cameroun	Egypte	Equateur
Egypte	Bolivie	Arabie Saoudite	Arabie Saoudite	Ghana	Afghanistan	Cote d' Ivoire
Bolivie	Algerie	Bresil	Cameroun	Bolivie	Finlande	Ghana
Cameroun	Costa Rica	Egypte	Afrique du sud	Bresil	Albanie	Colombie
Albanie	EAU	Guyana	Chine	Costa Rica	Cote d' Ivoire	Chine
Bulgarie	Haiti	Afrique du sud	EAU	Arabie Saoudite	Bulgarie	Hongrie
Equateur	Equateur	Algerie	Bolivie	Chine	Croatie	Afghanistan
Colombie	Egypte	Equateur	Bresil	Equateur	Guyana	Albanie
Algerie	Colombie	Colombie	Equateur	Algerie	France	Guyana
Costa Rica	Bresil	Argentine	Colombie	Colombie	Danemark	Costa Rica
Croatie	Chili	Albanie	Albanie	Chili	Argentine	Cameroun
Afrique du sud	Australie	Chine	Espagne	Bulgarie	Ghana	Chili
Bresil	Guyana	EAU	Chili	Argentine	Hongrie	Grece
Hongrie	Albanie	Bulgarie	Costa Rica	Egypte	Canada	Afrique du sud
Chili	Argentine	Coree du Sud	Grece	Afrique du sud	Grece	Bolivie
Arabie Saoudite	Chine	Chili	Chypre	Hongrie	Australie	Egypte
Grece	Coree du Sud	Hongrie	Argentine	Croatie	Allemagne	Coree du Sud
Coree du Sud	Chypre	Costa Rica	Coree du Sud	Australie	Colombie	EAU
Argentine	Canada	Croatie	Croatie	EAU	Equateur	Espagne
Chypre	Etats Unis	Chypre	Canada	Chypre	Arabie Saoudite	Chypre
Espagne	Grece	Grece	Etats Unis	Coree du Sud	Bolivie	Argentine
Australie	France	Etats Unis	Bulgarie	Etats Unis	Chili	Allemagne
Canada	Finlande	Espagne	Guyana	Grece	Etats Unis	Etats Unis
Finlande	Espagne	Canada	France	Allemagne	Bresil	Australie
France	Allemagne	France	Hongrie	Canada	Costa Rica	Finlande
EAU	Danemark	Australie	Allemagne	France	Chine	Danemark
Allemagne	Croatie	Allemagne	Danemark	Espagne	Coree du Sud	France
Etats Unis	Bulgarie	Finlande	Australie	Danemark	Chypre	Arabie Saoudite
Danemark	Hongrie	Danemark	Finlande	Finlande	EAU	Canada

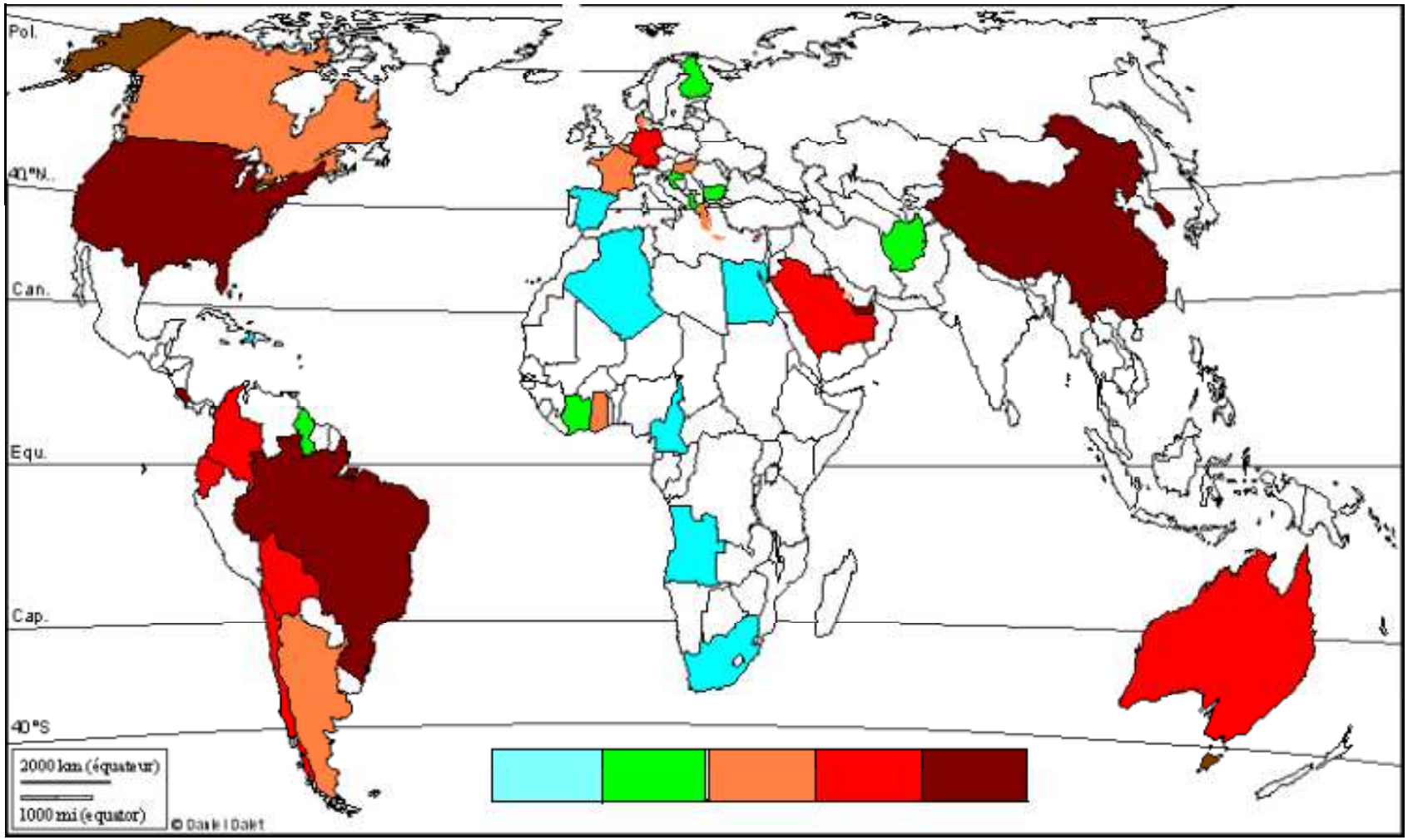
Regroupements

- Comment les regrouper, avec un niveau de regroupement variable ?
- Comment les classer ?
- Comment leur donner une note ?
- Impossibilité de classer selon tous les critères à la fois

Selon le produit national brut par habitant



Selon le taux de chômage



Du plus fort au plus faible taux de chômage

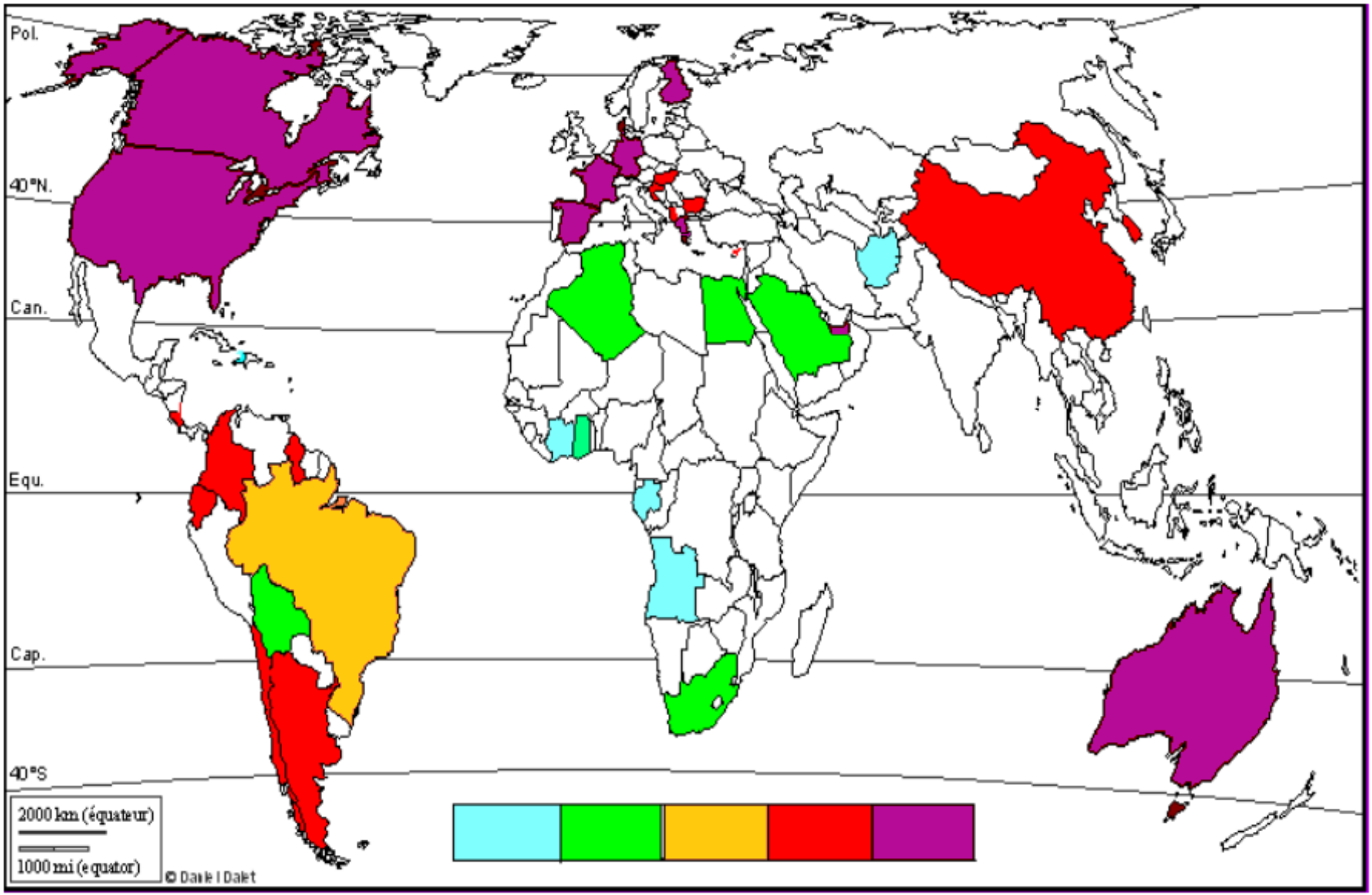
Carte de Kohonen

Allemagne Canada, France Danemark USA, Finlande	Australie Emirats	Afrique du Sud	Algérie Egypte Ghana
Espagne Grèce	Chypre Corée du Sud	Arabie saoudite Barhein	Cameroun Côte d'Ivoire
Bulgarie Croatie Hongrie	Argentine Chili		Afghanistan
Chine Equateur Colombie Costa-Rica	Albanie Guyana	Brésil	Angola

Carte de Kohonen

Allemagne Canada, France Danemark USA, Finlande	Australie Emirats	Afrique du Sud	Algérie Egypte Ghana
Espagne Grèce	Chypre Corée du Sud	Arabie saoudite Barhein	Cameroun Côte d'Ivoire
Bulgarie Croatie Hongrie	Argentine Chili		Afghanistan
Chine Equateur Colombie Costa-Rica	Albanie Guyana	Brésil	Angola

Selon la classification globale de Kohonen



Questions théoriques

[Cottrell and Fort, 1987, Cottrell et al., 1998, Fort, 2006]

- CM converge vers un minimum local de la distorsion
- VQ est un algorithme du gradient stochastique associé à la distorsion
- SOM n'est pas un algorithme du gradient dans le cas continu (espace des données connu par une densité)
- Dans le cas discret, SOM est un algorithme du gradient stochastique associé à la **Distorsion étendue**

$$E(m) = \frac{1}{2N} \sum_{i=1}^{N} \sum_{k=1}^{K} h_{kc(x_i)} \| m_k - x_i \|^2 \qquad (5)$$

E est une généralisation de la distorsion classique, mais n'est pas continue, et ne permet donc pas de montrer la convergence

- L'algorithme SOM Batch est une approximation d'un algorithme du gradient du second ordre associé à la Distorsion étendue

Cas le plus simple : $p = 1, \mathscr{X} = [0,1]$, réseau ficelle, densité uniforme

Théorème

Si ϵ constant <1/2 et si les voisinages sont $\{k-1, k, k+1\}$,

- *l'ensemble des suites monotones (croissantes ou décroissantes) est un ensemble absorbant atteint en un temps fini p.s.,*
- *le processus $m(t)$ converge en loi vers une loi stationnaire qui dépend de ϵ.*

Cas le plus simple : $p = 1, \mathscr{X} = [0,1]$, réseau ficelle, densité uniforme

Théorème

Si ϵ constant <1/2 et si les voisinages sont $\{k-1, k, k+1\}$,

- *l'ensemble des suites monotones (croissantes ou décroissantes) est un ensemble absorbant atteint en un temps fini p.s.,*
- *le processus $m(t)$ converge en loi vers une loi stationnaire qui dépend de ϵ.*

Théorème

Si ϵ tend vers 0 en vérifiant les conditions de Robbins-Monro

$$\sum_t \varepsilon(t) = +\infty \text{ and } \sum_t \varepsilon(t)^2 < +\infty. \tag{6}$$

une fois les prototypes ordonnés, le processus $m(t)$ converge p.s. vers une solution constante qu'on sait calculer.

Dimension 1, cas général, l'organisation

Théorème

On suppose la fonction de voisinage strictement décroissante à une certaine distance.

- *Les configurations croissantes ou décroissantes sont des états absorbants.*
- *Si ϵ est constant, le temps de mise en ordre est fini p.s..*

Dimension 1, cas général, la convergence

Théorème

On suppose la fonction de voisinage strictement décroissante à une certaine distance et la densité log-concave.

- *Si l'état initial est ordonné, il existe un point d'équilibre unique.*
- *Si ϵ est constant et l'état initial ordonné, alors il existe une loi invariante qui dépend de ϵ et qui converge en loi vers le point d'équilibre lorsque ϵ tend vers 0.*
- *Si $\epsilon(t)$ vérifie les conditions de Robbins-Monro (6) et si l'état initial est ordonné, alors le processus $m(t)$ converge p.s. vers cet équilibre unique.*

Ces résultats sont prouvés pour une fonction h qui ne dépend pas du temps, et on ne sait pas comment choisir $\epsilon(t)$ pour obtenir simultanément la mise en ordre et la convergence vers l'équilibre unique.

Dimension $p > 1$, espace continu

On ne peut pas définir d'ensemble absorbant. Les dispositions monotones sur chaque composantes ne sont pas absorbantes. Si on suppose $p = 2$ et si on note F^{++} l'ensemble des états des prototypes ayant leurs deux coordonnées croissantes, on a ces deux résultats apparemment contradictoires :

Théorème

- *Pour ϵ constant et des hypothèses très générales sur la densité f, le temps d'entrée dans F^{++} est fini avec une probabilité positive,*
- *mais dans le cas de 8 voisins, le temps de sortie est aussi fini avec une probabilité positive.*

Cependant, en pratique, l'algorithme converge bien vers un équilibre stable, sans qu'on sache précisément ni l'énoncer, ni le démontrer.

Données discrètes

[Ritter et al., 1992]

- Cas de toutes les applications en *data mining* et *clustering* de données
- Alors SOM est un algorithme du gradient stochastique associé à la distorsion étendue

$$E(m) = \frac{1}{2N} \sum_{i=1}^{N} \sum_{k=1}^{K} h_{kc(x_i)} \| m_k - x_i \|^2 \qquad (7)$$

- Cela ne permet pas de démontrer la convergence, mais la **distorsion étendue** est un critère de qualité
- Ce critère combine un critère de qualité du **Clustering** et un critère d'**Organisation** correcte
- Après apprentissage, un prototype doit être proche des données qu'il représente, mais aussi proche des prototypes voisins

La distorsion étendue

- La **distorsion étendue** est aussi utilisée comme critère dans le cas continu ou pour le SOM Batch
- Elle se calcule pour une fonction de voisinage donnée, indépendante du temps
- Dans la pratique, on commence les itérations avec un **voisinage large**, on fait décroître le rayon, et on termine **avec 0 voisin**
- SOM ou SOM Batch peuvent être utilisés comme initialisation de CM ou VQ, pour les accélérer et obtenir de meilleurs minima locaux

Variantes de SOM

De nombreuses variantes de SOM pour données numériques ont été définies

- Modification du calcul de l'unité gagnante [Heskes, 1999] pour que SOM dans le cas continu soit un algorithme de gradient
- Remplacement du calcul de l'unité gagnante par un tirage probabiliste [Graepel et al., 1998]
- Définition de modèles de mélange avec des contraintes de voisinages, qui reproduisent certaines propriétés des SOM et SOM Batch
- ...

Extensions pour traiter des données non numériques

Diapos précédentes : les données sont décrites par des variables numériques

Les cartes auto-organisées ont été étendues au cas :

- de données de sondages (les variables sont qualitatives, comme dans le cas de données de sondages dans lesquelles les variables sont des réponses à des questions à choix multiples) ;

Extensions pour traiter des données non numériques

Diapos précédentes : les données sont décrites par des variables numériques

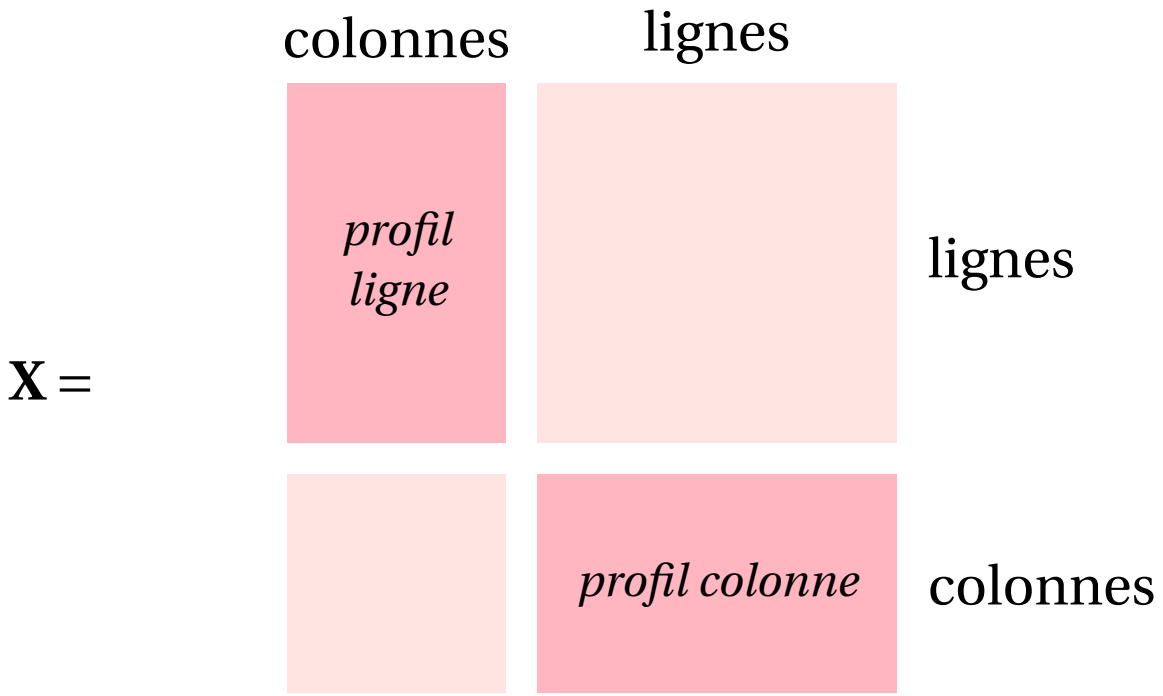
Les cartes auto-organisées ont été étendues au cas :

- de données de sondages (les variables sont qualitatives, comme dans le cas de données de sondages dans lesquelles les variables sont des réponses à des questions à choix multiples) ;
- de données décrites par une matrice de dissimilarité ou un noyau (les observations sont décrites par leurs relations deux à deux) : graphes, séries temporelles qualitatives ...

Tables de contingence

KORRESP [Cottrell et al., 1993]

Les données : une table de contingence (deux variables qualitatives) $\mathbf{T} = (n_{ij})_{ij}$ avec p lignes et q colonnes. $\Rightarrow$ création des données numériques $\mathbf{X}$:



où

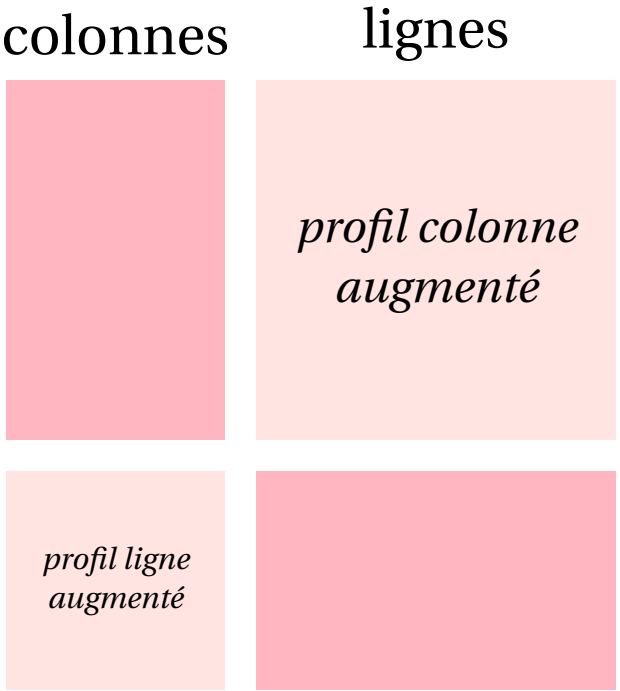
- $\forall\, i = 1, \ldots, p$ et $\forall\, j = 1, \ldots, q$, $\mathbf{x}_{ij} = \frac{n_{ij}}{\sqrt{n_{i.}}} \times \sqrt{\frac{n}{n_{.j}}}$

Tables de contingence

KORRESP [Cottrell et al., 1993]

Les données : une table de contingence (deux variables qualitatives) $\mathbf{T} = (n_{ij})_{ij}$ avec p lignes et q colonnes. $\Rightarrow$ création des données numériques $\mathbf{X}$:

$\mathbf{X} =$



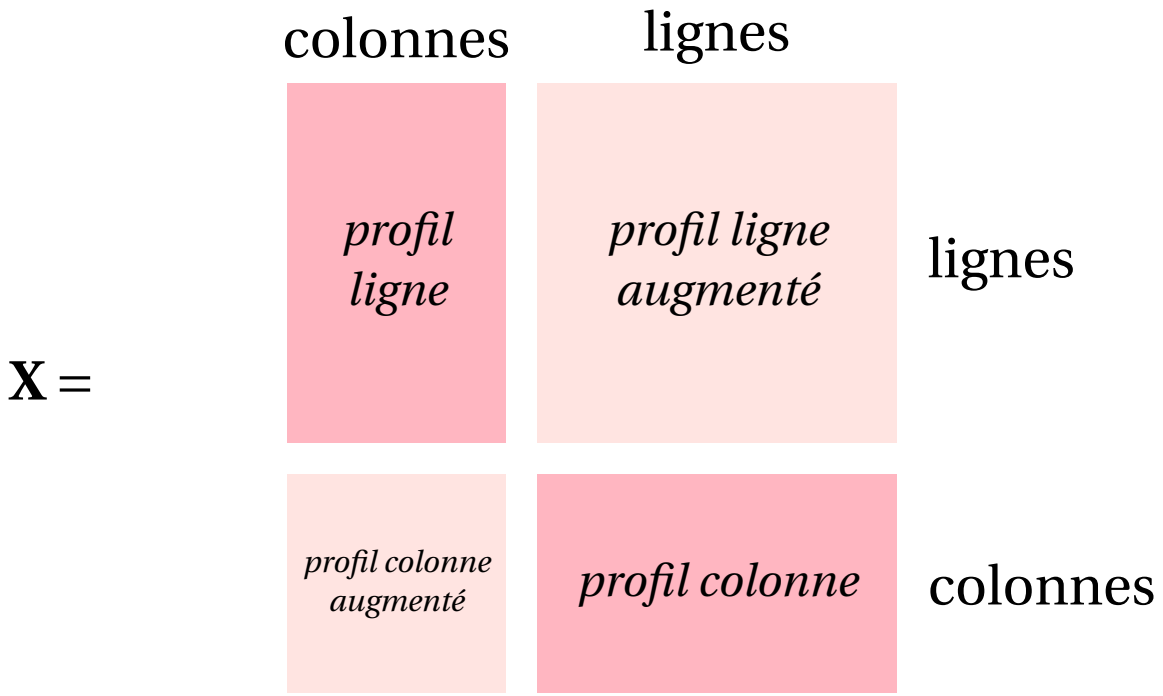
où

- $\forall\, i = 1, \ldots, p$ et $\forall\, j = q+1, \ldots, q+p$, $\mathbf{x}_{ij} = \mathbf{x}_{k(i)+p,j}$ avec $k(i) = \arg\max_{k=1,\ldots,q} \mathbf{x}_{ik}$

Tables de contingence

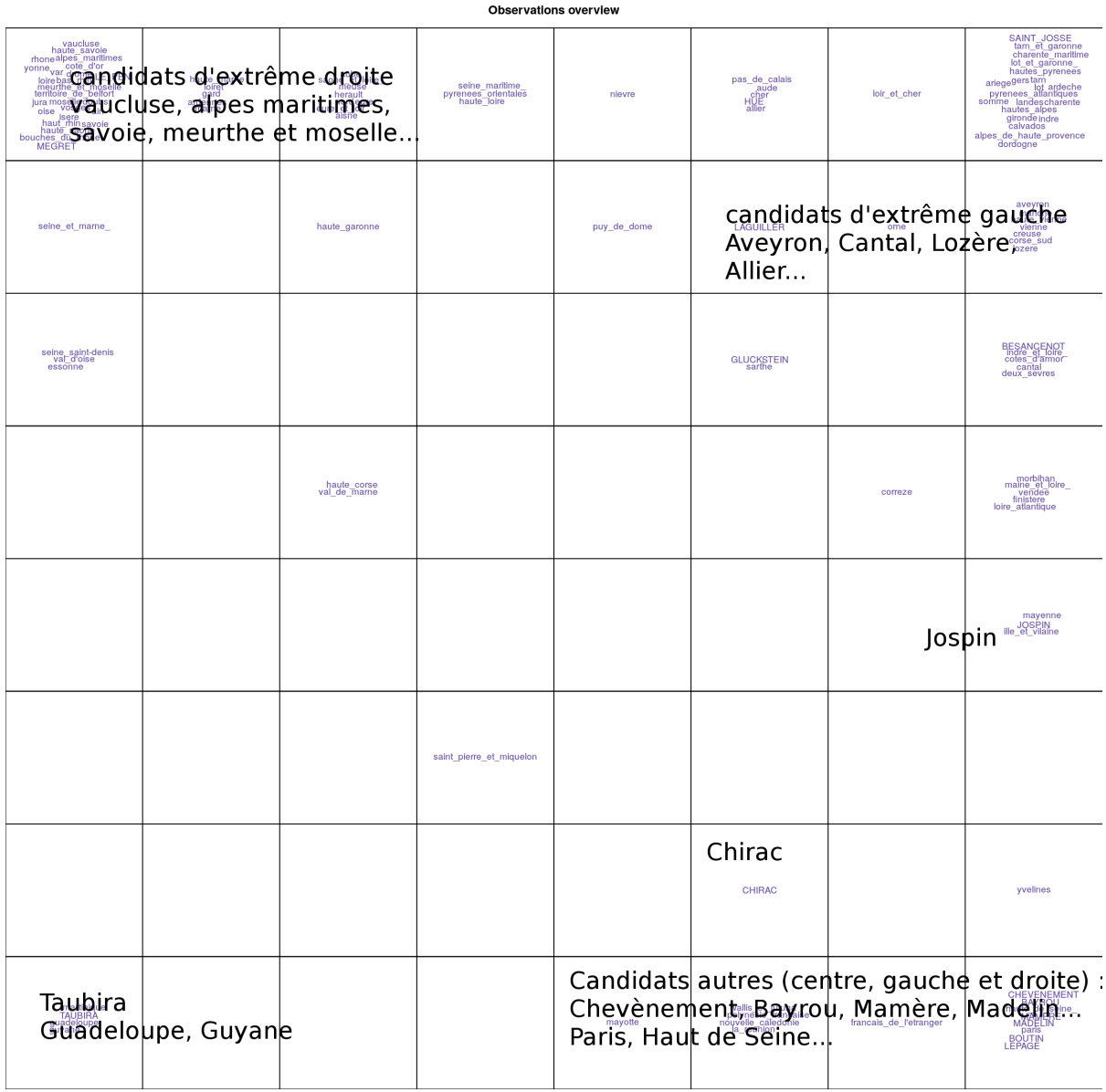
KORRESP [Cottrell et al., 1993]

Les données : une table de contingence (deux variables qualitatives) $\mathbf{T} = (n_{ij})_{ij}$ avec p lignes et q colonnes. $\Rightarrow$ création des données numériques $\mathbf{X}$:



- affectation utilise les profils réduits
- mise à jour des prototypes utilise les profils augmentés
- traite alternativement une ligne et une colonne

Exemple : les élections présidentielles de 2002



... pour un nombre quelconque de variables qualitatives

Principe général de KORRESP : calcul de profils pour lesquels la distance Euclidienne est la distance du χ^2 utilisée entre profils lignes (ou profils colonnes) dans une AFC

... pour un nombre quelconque de variables qualitatives

Principe général de KORRESP : calcul de profils pour lesquels la distance Euclidienne est la distance du χ^2 utilisée entre profils lignes (ou profils colonnes) dans une AFC

De même, pour un nombre quelconque de variables qualitatives : transformation des données de sorte que la distance euclidienne entre observations soit la même que la distance utilisée dans une AFCM
[Cottrell and Letrémy, 2005]

Données décrites par des similarités / dissimilarités

SOM "relationnel" [Hammer and Hasenfuss, 2010, Olteanu and Villa-Vialaneix, 2015a]

Les données : sont décrites par une matrice de (dis)similarité $\mathbf{D} = (\delta(x_i, x_j))_{i,j=1,\dots,n}$ ($(x_i)_i$ ne prennent pas nécessairement des valeurs dans un espace vectoriel)

Données décrites par des similarités / dissimilarités

SOM "relationnel" [Hammer and Hasenfuss, 2010, Olteanu and Villa-Vialaneix, 2015a]

Les données : sont décrites par une matrice de (dis)similarité $\mathbf{D} = (\delta(x_i, x_j))_{i,j=1,\dots,n}$ ($(x_i)_i$ ne prennent pas nécessairement des valeurs dans un espace vectoriel)

[Goldfarb, 1984] : contexte pseudo-euclidien : $\exists$ deux espaces euclidiens $\mathcal{E}_1$ et $\mathcal{E}_2$ et $\psi_1 : \{x_i\} \to \mathcal{E}_1$, $\psi_2 : \{x_i\} \to \mathcal{E}_2$ tels que :

$$\delta(x_i, x_j) = \|\psi_1(x_i) - \psi_1(x_j)\|^2_{\mathcal{E}_1} - \|\psi_2(x_i) - \psi_2(x_j)\|^2_{\mathcal{E}_2}$$

Données décrites par des similarités / dissimilarités

SOM "relationnel" [Hammer and Hasenfuss, 2010, Olteanu and Villa-Vialaneix, 2015a]

Principe : utiliser la représentation des données dans $\mathcal{E} = \mathcal{E}_1 \otimes \mathcal{E}_2$.

- prototypes : sont exprimés comme des combinaisons convexes (symboliques) des $(x_i)_i$: $p_u \sim \sum_{i=1}^{n} \gamma_{ui} x_i$, $\gamma_{ui} \geq 0$ et $\sum_i \gamma_{ui} = 1$ (en fait des $(\psi_1(x_i)), \psi_2(x))$)

Données décrites par des similarités / dissimilarités

SOM "relationnel" [Hammer and Hasenfuss, 2010, Olteanu and Villa-Vialaneix, 2015a]

Principe : utiliser la représentation des données dans $\mathcal{E} = \mathcal{E}_1 \otimes \mathcal{E}_2$.

- prototypes : sont exprimés comme des combinaisons convexes (symboliques) des $(x_i)_i$: $p_u \sim \sum_{i=1}^n \gamma_{ui} x_i$, $\gamma_{ui} \geq 0$ et $\sum_i \gamma_{ui} = 1$ (en fait des $(\psi_1(x_i)), \psi_2(x))$)
- calcul de la distance : $\|x_i - p_u\|^2_{\mathcal{E}}$ est

$$\left(\mathbf{D}\gamma_u\right)_i - \frac{1}{2}\gamma_u^T \mathbf{D}\gamma_u$$

Données décrites par des similarités / dissimilarités

SOM "relationnel" [Hammer and Hasenfuss, 2010, Olteanu and Villa-Vialaneix, 2015a]

Principe : utiliser la représentation des données dans $\mathcal{E} = \mathcal{E}_1 \otimes \mathcal{E}_2$.

- prototypes : sont exprimés comme des combinaisons convexes (symboliques) des $(x_i)_i$: $p_u \sim \sum_{i=1}^n \gamma_{ui} x_i$, $\gamma_{ui} \geq 0$ et $\sum_i \gamma_{ui} = 1$ (en fait des $(\psi_1(x_i)), \psi_2(x))$)
- calcul de la distance : $\|x_i - p_u\|_{\mathcal{E}}^2$ est

$$\left(\mathbf{D}\gamma_u\right)_i - \frac{1}{2}\gamma_u^T \mathbf{D}\gamma_u$$

- mise à jour des prototypes : par une mise à jour de leurs coordonnées $(\gamma_u)_u$:

$$\gamma_u^{t+1} \leftarrow \gamma_u^t + \mu(t) H^t(d(\mathscr{C}(x_i), u)) \left(\mathbf{1}_i - \gamma_u^t\right)$$

avec $\mathbf{1}_{il} = 1$ if $l = i$ et 0 sinon.

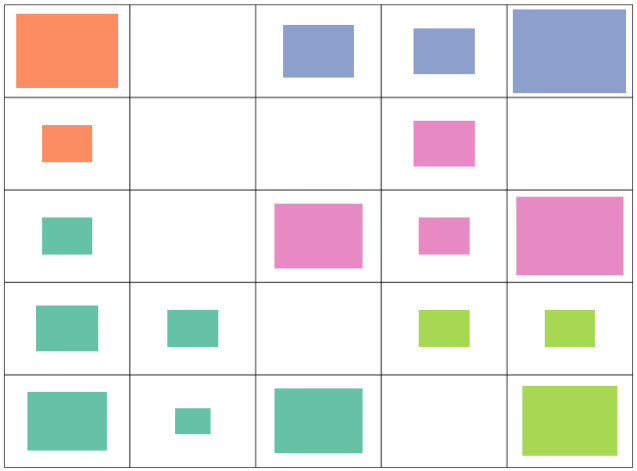
Exemple 1 : un graphe

Le graphe : graphe de co-apparition (dans un même chapitre) des personnages du roman “Les misérables”

Dissimilarité : longueur du plus court chemin entre deux sommets

Principe : [Olteanu and Villa-Vialaneix, 2015b]

- SOM relationnel
- classification ascendante hiérarchique des prototypes pour déterminer des “super-classes”



Exemple 1 : un graphe

Le graphe : graphe de co-apparition (dans un même chapitre) des personnages du roman “Les misérables”

Dissimilarité : longueur du plus court chemin entre deux sommets

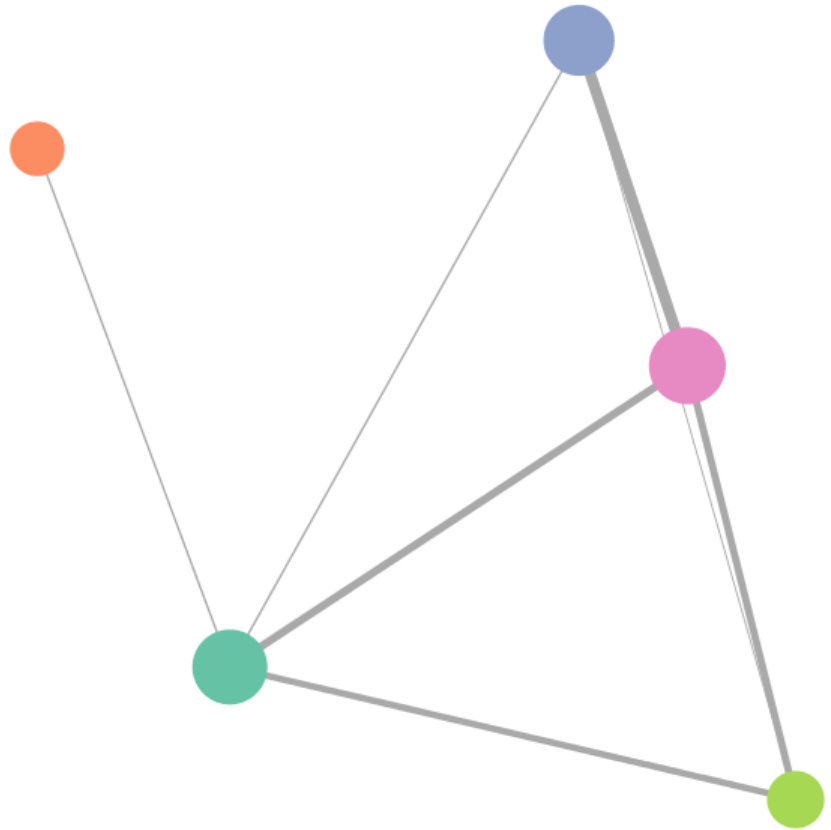
Principe : [Olteanu and Villa-Vialaneix, 2015b]

- SOM relationnel
- classification ascendante hiérarchique des prototypes pour déterminer des “super-classes”
- projection du graphe : chaque super-classe est représentée par un disque de rayon proportionnel au nombre de sommets dans la super-classe et positionné au centre de gravité de la super-classe sur la grille. Les arêtes ont une épaisseur proportionnelle au nombre de connexions entre deux classes

Marie Cottrell[1] &
Nathalie
Villa-Vialaneix[2]

Exemple 1 : un graphe

Le graphe : graphe de co-apparition (dans un même chapitre) des personnages du roman “Les misérables”
Dissimilarité : longueur du plus court chemin entre deux sommets



Exemple 2 : trajectoire d'emplois

Les données : "Génération 98" à 7 ans - 2005, CEREQ, Centre Maurice Halbwachs (CMH). 16 040 personnes ayant obtenu leur diplôme de fin d'études en 1998 et suivis pendant 94 mois après l'école. Chaque mois, enregistrement de leur activité (CDI, CDD, apprentissage, fonctionnaire...)

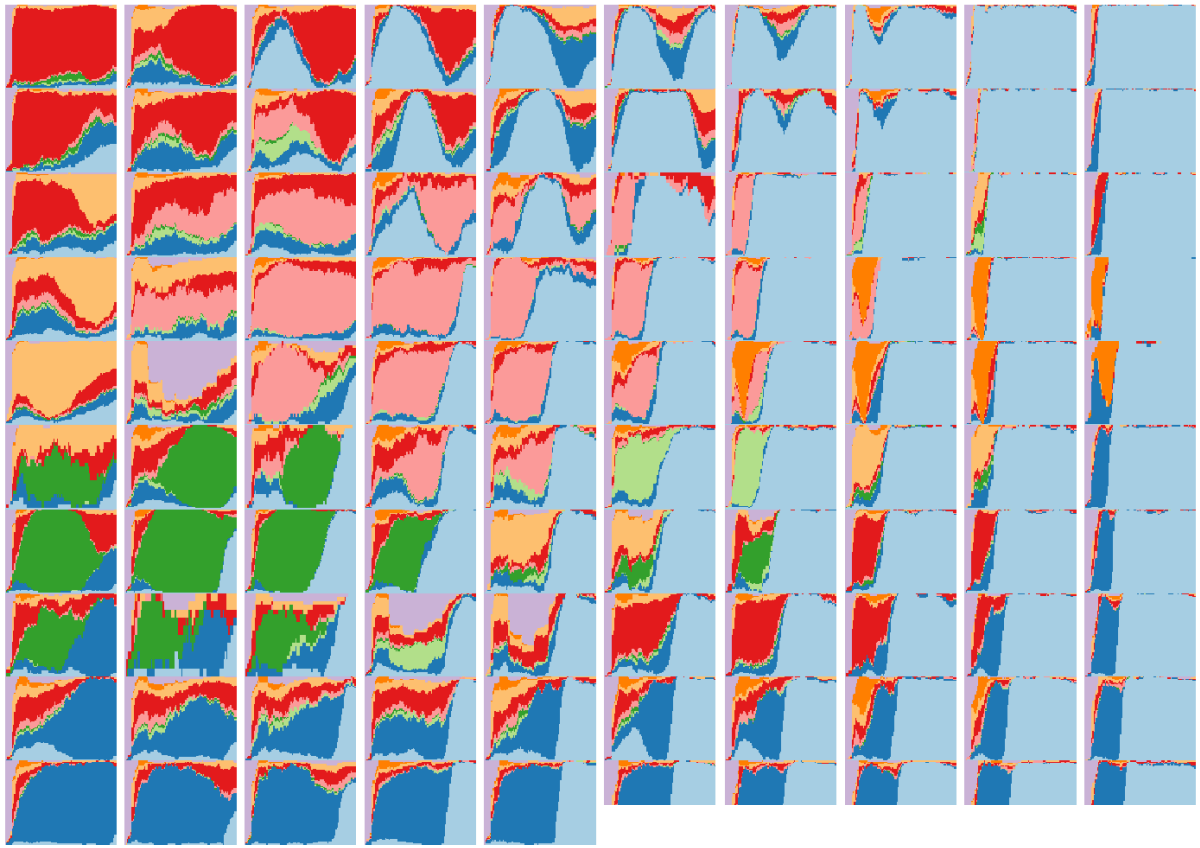
Exemple 2 : trajectoire d'emplois

Les données : "Génération 98" à 7 ans - 2005, CEREQ, Centre Maurice Halbwachs (CMH). 16 040 personnes ayant obtenu leur diplôme de fin d'études en 1998 et suivis pendant 94 mois après l'école. Chaque mois, enregistrement de leur activité (CDI, CDD, apprentissage, fonctionnaire...)

Dissimilarité : combinaison linéaire (optimisée par une descente de gradient stochastique au cours de l'apprentissage : voir [Olteanu and Villa-Vialaneix, 2015a] pour les détails) de plusieurs distances entre séries catégorielles
[Needleman and Wunsch, 1970, Abbott and Forrest, 1986] (type distances d'édition

Marie Cottrell[1] &
Nathalie
Villa-Vialaneix[2]

Exemple 2 : trajectoire d'emplois



haut gauche : exclusion
droite : insertion rapide

Stochasticité des résultats

Constat : plusieurs mises en œuvre de l'algorithme SOM stochastique donnent des résultats différents (même avec la même initialisation).

Stochasticité des résultats

Constat : plusieurs mises en œuvre de l'algorithme SOM stochastique donnent des résultats différents (même avec la même initialisation).

Améliorer la stabilité (SOM ensemble)
[Petrakieva and Fyfe, 2003, Saavedra et al., 2007, Vrusias et al., 2007, Baruque and Corchado, 2011, Mariette et al., 2014, Mariette and Villa-Vialaneix, 2016]

Stochasticité des résultats

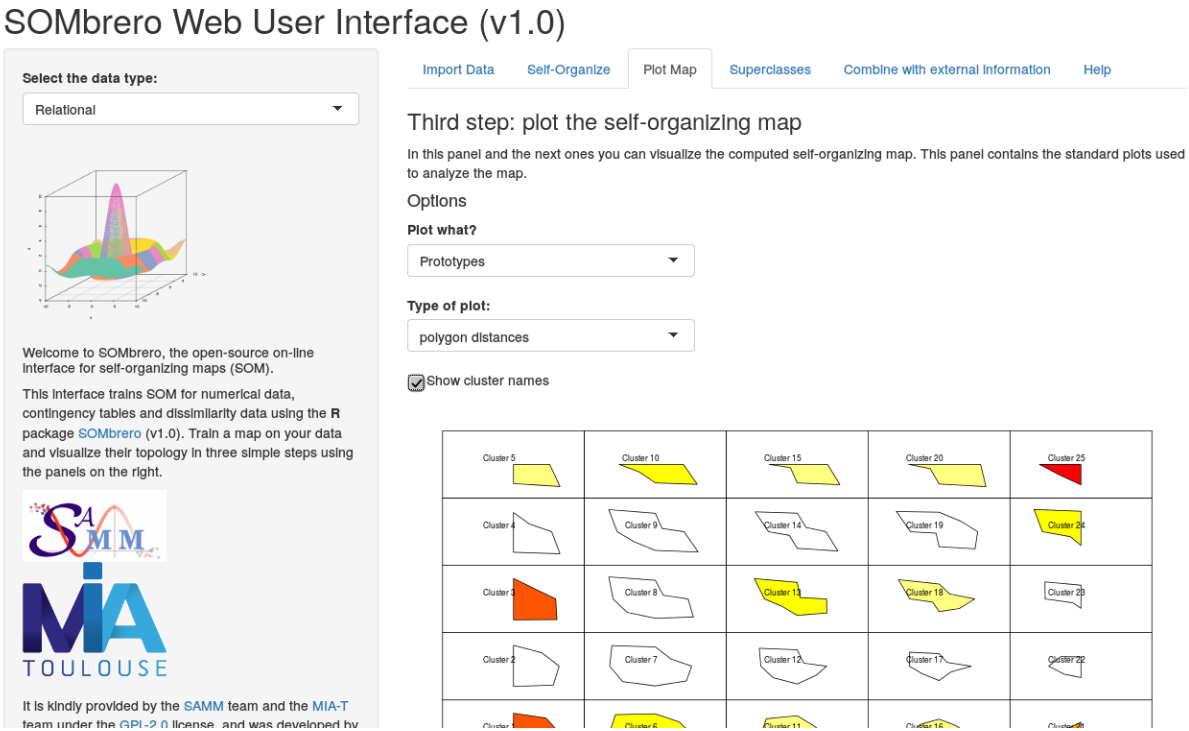
Constat : plusieurs mises en œuvre de l'algorithme SOM stochastique donnent des résultats différents (même avec la même initialisation).

Améliorer la stabilité (SOM ensemble)
[Petrakieva and Fyfe, 2003, Saavedra et al., 2007, Vrusias et al., 2007, Baruque and Corchado, 2011, Mariette et al., 2014, Mariette and Villa-Vialaneix, 2016]

Utiliser la stochasticité pour évaluer la qualité des résultats avec des indices de stabilité [de Bodt et al., 2002]. Cette approche est utilisée pour de la fouille de textes (médiévaux) dans [Bourgeois et al., 2015]

SOM en pratique...

- batch SOM pour données numériques et données relationnelles est implémenté dans **yasomi** `http://yasomi.r-forge.r-project.org`
- KORRESP et SOM stochastique pour données numériques et relationnelles sont implémentés dans **SOMbrero** (CRAN)



Conclusion

SOM et SOM Batch sont des algorithmes de *Clustering* qui ont des propriétés très intéressantes

- *Complexité linéaire* en le nombre de données, adapté aux données très nombreuses (*big data*)
- Propriétés de *visualisation* des données et des classes
- Utilisation avec des données incomplètes, *estimation des données manquantes*
- Bonne initialisation et *accélération* des algorithmes à 0 voisin

Conclusion

SOM et SOM Batch sont des algorithmes de *Clustering* qui ont des propriétés très intéressantes

- *Complexité linéaire* en le nombre de données, adapté aux données très nombreuses (*big data*)
- Propriétés de *visualisation* des données et des classes
- Utilisation avec des données incomplètes, *estimation des données manquantes*
- Bonne initialisation et *accélération* des algorithmes à 0 voisin

La version relationnelle donne une alternative intéressante pour des données non numériques mais sa complexité est accrue et l'interprétabilité diminuée (représentation des résultats, interprétation des prototypes)

Abbott, A. and Forrest, J. (1986).

Optimal matching methods for historical sequences.
Journal of Interdisciplinary History, 16 :471–494.

Baruque, B. and Corchado, E. (2011).

Fusion methods for unsupervised learning ensembles, volume 322 of *Studies in Computational Intelligence*.
Springer.

Bourgeois, N., Cottrell, M., Deruelle, B., Lamassé, S., and Letrémy, P. (2015).

How to improve robustness in Kohonen maps and display additional information in factorial analysis : application to text mining.
Neurocomputing, 147 :120–135.

Cottrell, M. and Fort, J. (1987).

Étude d'un processus d'auto-organisation.
Annales de l'IHP, section B, 23(1) :1–20.

Cottrell, M., Fort, J., and Pagès, G. (1998).

Theoretical aspects of the SOM algorithm.
Neurocomputing, 21 :119–138.

Cottrell, M. and Letrémy, P. (2005).

How to use the Kohonen algorithm to simultaneously analyse individuals in a survey.
Neurocomputing, 63 :193–207.

Cottrell, M., Letrémy, P., and Roy, E. (1993).

Analyzing a contingency table with Kohonen maps : a factorial correspondence analysis.
In Cabestany, J., Mary, J., and Prieto, A. E., editors, *Proceedings of International Workshop on Artificial Neural Networks (IWANN 93)*, Lecture Notes in Computer Science, pages 305–311. Springer Verlag.

de Bodt, E., Cottrell, M., and Verleisen, M. (2002).

Statistical tools to assess the reliability of self-organizing maps.
Neural Networks, 15(8-9) :967–978.


Fort, J. (2006).
SOM's mathematics.
Neural Networks, 19(6-7) :812–816.


Goldfarb, L. (1984).
A unified approach to pattern recognition.
Pattern Recognition, 17(5) :575–582.


Graepel, T., Burger, M., and Obermayer, K. (1998).
Self-organizing maps : generalizations and new optimization techniques.
Neurocomputing, 21 :173–190.


Hammer, B. and Hasenfuss, A. (2010).
Topographic mapping of large dissimilarity data sets.
Neural Computation, 22(9) :2229–2284.


Heskes, T. (1999).
Energy functions for self-organizing maps.
In Oja, E. and Kaski, S., editors, *Kohonen Maps*, pages 303–315. Elsevier, Amsterdam.


Kohonen, T. (1982a).
Analysis of a simple self-organizing process.
Biological Cybernectis, 44 :135–140.


Kohonen, T. (1982b).
Self-organized formation of topologically correct feature maps.
Biological Cybernetics, 43 :59–69.


Kohonen, T. (1995).
Self-Organizing Maps, volume 30 of *Springer Series in Information Science*.


Mariette, J., Olteanu, M., Boelaert, J., and Villa-Vialaneix, N. (2014).
Bagged kernel SOM.

In Villmann, T., Schleif, F., Kaden, M., and Lange, M., editors, *Advances in Self-Organizing Maps and Learning Vector Quantization (Proceedings of WSOM 2014)*, volume 295 of *Advances in Intelligent Systems and Computing*, pages 45–54, Mittweida, Germany. Springer Verlag, Berlin, Heidelberg.

Mariette, J. and Villa-Vialaneix, N. (2016).

Aggregating self-organizing maps with topology preservation.
In *Proceedings of WSOM 2016*, Houston, TX, USA.
Forthcoming.

Needleman, S. and Wunsch, C. (1970).

A general method applicable to the search for similarities in the amino acid sequence of two proteins.
Journal of Molecular Biology, 48(3) :443–453.

Olteanu, M. and Villa-Vialaneix, N. (2015a).

On-line relational and multiple relational SOM.
Neurocomputing, 147 :15–30.

Olteanu, M. and Villa-Vialaneix, N. (2015b).

Using SOMbrero for clustering and visualizing graphs.
Journal de la Société Française de Statistique.
Forthcoming.

Petrakieva, L. and Fyfe, C. (2003).

Bagging and bumping self organising maps.
Computing and Information Systems Journal, 9 :69–77.

Ritter, H., Martinetz, T., and Schulten, K. (1992).

Neural Computation and Self-Organizing Maps : an Introduction.
Addison-Wesley.

Saavedra, C., Salas, R., Moreno, S., and Allende, H. (2007).

Fusion of self organizing maps.
In *Proceedings of the 9th International Work-Conference on Artificial Neural Networks (IWANN 2007)*.

Vrusias, B., Vomvoridis, L., and Gillam, L. (2007).

Distributing SOM ensemble training using grid middleware.
In *Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN 2007)*, pages 2712–2717.