

ANNEXE VI.1

**Classification
multiclasse non exhaustive**

Etude générale,
proposition fondée sur les Nuées Dynamiques,
réalisation d'un logiciel

Table des matières de l'Annexe VI.1

A. EXPOSÉ DE LA PROBLÉMATIQUE.....	683
1. Contexte initiateur	683
a) <i>Le regroupement d'unités linguistiques.....</i>	683
b) <i>L'intérêt du multiclassement.....</i>	683
c) <i>L'intérêt d'un classement non exhaustif.....</i>	683
2. Les algorithmes existants.....	683
a) <i>Nature de la classification obtenue</i>	683
b) <i>Rigueur théorique</i>	684
3. Les Nuées Dynamiques	684
a) <i>Origine et positionnement.....</i>	684
b) <i>Description de l'algorithme des nuées dynamiques</i>	684
c) <i>Comparaison avec les méthodes de la même famille</i>	685
Les centres mobiles, de Forgy	685
La méthode de Jancey.....	685
Les <i>k-means</i> de Mac Queen.....	686
d) <i>Atouts</i>	686
e) <i>Points faibles.....</i>	687
4. Mention des méthodes alternatives envisagées les plus intéressantes	687
a) <i>L'algorithme de Van Den Driessche</i>	687
Présentation.....	687
Discussion	688
b) <i>Deux classifications ascendantes hiérarchiques expérimentalement éprouvées et appréciées.....</i>	688
Rappel : raisons du rejet des CAH.....	688
Utiliser une CAH.....	688
Méthode de Ward & distance du χ^2	689
Average & indice de Jaccard	689
c) <i>Point sur les méthodes disponibles et / ou pratiquées dans l'équipe</i>	689
Outils.....	689
Conception d'une tactique basée sur les Nuées Dynamiques	689
B. PRÉSENTATION DU PROGRAMME	690
1. Conventions	690
a) <i>Vocabulaire.....</i>	690
b) <i>Notations.....</i>	690
2. Entrées / sorties	690
a) <i>Paramètres de la ligne de commande.....</i>	690
b) <i>Format et contraintes sur le fichier de données</i>	691
c) <i>Format du fichier résultats</i>	691

3. Disponibilité.....	693
C. ALGORITHME	694
1. Ressources.....	694
a) <i>Constantes et variables globales</i>	694
b) <i>Fonctions.....</i>	694
2. Traitement	695
a) <i>Série initiale de classifications.....</i>	695
b) <i>Détermination des formes fortes.....</i>	697
c) <i>Degré de variation des formes fortes.....</i>	697
d) <i>Regroupement des formes fortes.....</i>	698
e) <i>Classement des individus restants</i>	698
D. DISCUSSION	700
1. Explication des apports à l'algorithme original.....	700
a) <i>Equilibre général.....</i>	700
b) <i>La fonction diamètre D et la fonction poids P.....</i>	700
c) <i>La classe Z des non affectés.....</i>	701
d) <i>Degré de variation des formes fortes.....</i>	701
e) <i>Regroupement des formes fortes.....</i>	702
f) <i>Classement final.....</i>	703
2. Points d'évolution.....	703
a) <i>Taille des noyaux</i>	703
b) <i>Distance D</i>	704
c) <i>L'agrégation-écartement R.....</i>	704
Fonctions proposées par Diday	704
Propositions pour une autre fonction R	705
d) <i>La fonction de représentation.....</i>	705
e) <i>Ex-æquo</i>	706
f) <i>Génération d'un autre noyau.....</i>	706
g) <i>Gestion de la classe Z.....</i>	706
h) <i>Convergence</i>	707
Théorie et pratique	707
Heuristiques pour garantir la robustesse du traitement et garder des résultats acceptables	707
E. APPENDICE	709
1. Préparation des données : choix de l'indice de dissimilarité.....	709
Décrire les indices existants	709
Interpréter les indices	709
Sélection d'un indice.....	709

F. REPÈRES BIBLIOGRAPHIQUES	711
1. Article de référence sur la méthode.....	711
2. Présentation et commentaires concernant les Nuées Dynamiques	711
3. Documents complémentaires : classifications en général ; prolongements possibles	711

A. EXPOSÉ DE LA PROBLÉMATIQUE

1. Contexte initiateur

a) *Le regroupement d'unités linguistiques*

Dans le cadre de la réalisation d'un outil de diffusion ciblée de documents¹, on souhaite la construction automatique d'unités descriptives, regroupant des unités élémentaires trouvées dans les textes d'un corpus.

En effet, la manifestation dans un texte de certains groupements d'unités apporte une information sémantique, par exemple sur le thème abordé dans le texte. La définition de ces unités d'ordre supérieur est décisif pour la qualité de la représentation des textes. Une aide automatisée à leur construction est nécessaire pour leur donner une envergure opérationnelle.

b) *L'intérêt du multiclassement*

Une unité linguistique élémentaire peut contribuer de plusieurs manières à la sémantique du texte. Les unités sémantiques complexes que l'on veut former sont distinctes, en revanche elles peuvent reposer sur certains éléments communs.

La problématique ne se pose pas en termes de *répartition* des unités élémentaires dans des classes d'équivalence. Les unités sont plutôt le matériau qui sert à la formation de regroupements significatifs. Un même composant peut être utile à plusieurs groupements.

La formation de classes non disjointes rend compte de la non univocité des unités à regrouper, par rapport à la description donnée par les regroupements. C'est normalement le cas quand il s'agit de deux plans différents. Dans l'application présentée par exemple, le plan de l'expression (réalisée matériellement par les mots, les lettres, etc.) n'est pas en correspondance univoque avec le plan du contenu sémantique : il y a plusieurs manières d'exprimer une même idée, un même mot peut contribuer à exprimer différentes idées.

c) *L'intérêt d'un classement non exhaustif*

Il n'y a pas de raison que toute unité du texte soit utilisée. S'obliger à utiliser toute unité, c'est courir le risque d'avoir des regroupements moins cohérents, et des affectations non significatives d'unités à des regroupements.

Il n'est pas rare que les données ne se prêtent pas à une classification exhaustive : les partitions comportent quelquefois dans ce cas une classe « divers », qui n'a ni la cohérence ni la régularité des autres classes. L'intérêt d'une classification non exhaustive est de ne pas définir cette classe non motivée, qui n'est en fait qu'un artefact.

2. Les algorithmes existants

a) *Nature de la classification obtenue*

Les algorithmes développés en analyse des données fournissent soit une classification hiérarchique, soit une partition.

Une partition est une classification qui attribue chaque élément à une classe et une seule. La classification n'est donc pas multiclasse, les classes sont disjointes, sans recouvrement. La classification est exhaustive : tout élément, sans exception, est *in fine* intégré à une classe.

¹ Il s'agit de DECID, le serveur de *Diffusion Electronique Ciblée d'Informations et de Documents*, mis à disposition sur l'intranet EDF. Il a été conçu au sein du Département SID (*Systèmes d'Information et de Documentation*) de la Direction des Etudes et Recherches d'EDF.

Une classification hiérarchique se traduit par une suite de classifications s'emboîtant les unes dans les autres. Pour chacune de ces classifications, un élément est affecté à une seule classe ou à aucune. La classification n'est donc jamais multiclasse. La suite de classifications comporte par construction des classifications non exhaustives. La non affectation d'un élément à une classe ne devient significative qu'en présence de classes assez consistantes, auxquelles l'élément n'est pas rattaché. La classification doit être à un stade où le seuil de liaison des classes est intermédiaire, ni trop faible (classes lâches et attachement abusif d'éléments), ni trop fort (des éléments ne sont pas intégrés à une classe malgré leur proximité à celle-ci). La difficulté d'exploiter une classification hiérarchique pour obtenir un ensemble de classes de « même niveau » est justement de se donner un critère satisfaisant et efficace pour choisir une des classifications parmi toutes celles de la série.

b) Rigueur théorique

La construction de certaines classifications est une mécanique huilée, fondée sur un certain nombre de résultats mathématiques précis. Les possibilités de réaménagements sont réduites : une modification peut remettre en cause tout l'appareil théorique qui fonde l'algorithme. Les théories sous-jacentes étant souvent très complexes, il est difficile de prévoir l'impact d'une modification. Ces méthodes s'avèrent donc rigides.

Les modifications semblent plutôt pouvoir se faire pour les classifications correspondant à une approche heuristique, par nature assez robuste. Il est aussi tout à fait possible de faire des développements en amont ou en aval des procédures de classification : préparer les données et les présenter sous une certaine forme, exploiter toutes les informations contenues dans les résultats. Ajouter d'autres traitements est d'autant plus envisageable que la classification sur laquelle on se greffe a de bonnes performances.

3. Les Nuées Dynamiques

a) Origine et positionnement

Le nom associé à la méthode des nuées dynamiques est celui d'E. Diday.

La proposition de Diday s'inscrit dans le contexte d'autres méthodes comparables. Elle peut être interprétée comme une extension de la méthode des centres mobiles (de Forgy). La méthode de Jancey, et la méthode des *k-means* de Mac Queen sont d'autres variantes de la même famille, apparues également à la fin des années soixante. Il est intéressant de considérer l'ensemble de ces méthodes pour mieux percevoir les lignes de force de la démarche, et les choix faits dans chacune avec leur signification.

b) Description de l'algorithme des nuées dynamiques

La présentation que donne Michel Volle des Nuées dynamiques (Volle 1985, §XVII) convient parfaitement pour expliquer les principes de la méthode et introduire les concepts qui lui sont propres. Nous reprenons donc ses mots dans les paragraphes qui suivent.

Le nom fort bien choisi qu'a reçu cette méthode lui confère d'emblée un certain prestige auprès de ceux qui, ne connaissant pas l'analyse des données, éprouvent devant cette dénomination poétique l'impression agréable (mais trompeuse) d'être au bord de la compréhension intuitive d'un mystère. Il ne s'agit cependant que d'un outil logique comparable aux autres.

Supposons donné un ensemble E sur lequel sont définies une distance $d(x,y)$ entre éléments et une distance $D(X,Y)$ entre sous-ensembles. Prenons au hasard k paquets de p points chacun dans E ; nous appellerons chacun de ces paquets un *noyau*. Ces noyaux nous permettent de définir une partition de E en k classes, chaque classe comprenant les éléments qui sont plus proches d'un des noyaux que de tous les autres noyaux. A partir de cette partition, on définit une nouvelle famille de noyaux en associant à chaque classe de la partition l'ensemble de p points qui en est le plus proche. Puis on recommence : à cette nouvelle famille de noyaux va être associée une nouvelle partition, etc. Il est facile de démontrer que le procédé converge sous certaines conditions : on finit par aboutir à une partition et à une famille de noyaux qui se correspondent.

Ainsi la méthode des nuées dynamiques permet de construire par itération, à partir d'une famille absolument quelconque de noyaux, une partition en k classes. Mais cette construction contient une part d'arbitraire : la partition obtenue dépend du choix initial des noyaux. Pour compenser dans une certaine mesure cet arbitraire, on applique la méthode plusieurs fois de suite, en partant à chaque fois d'une famille différente de noyaux tirée au hasard. On obtient ainsi plusieurs partitions en k classes. Si l'on considère deux éléments quelconques, ils peuvent avoir été classés ensemble dans certaines de ces partitions, et classés séparément dans d'autres. On appelle *forme forte* –encore une dénomination bien trouvée– un ensemble d'éléments qui auront été classés ensemble dans toutes les classifications. Ces formes fortes déterminent une partition de E qui peut fort bien contenir plus de k classes. Les formes fortes qui ne comprennent qu'un élément ne présentent pas d'intérêt, car elles concernent des éléments qui semblent « inclassables ». Par contre, les formes fortes qui comprennent beaucoup d'éléments sont intéressantes : pour qu'un groupe d'éléments ait résisté aux aléas dus aux choix des différentes familles de noyaux, il fallait qu'il fût bien homogène.

Il est intéressant de noter que l'algorithme est entièrement décrit par la donnée de :

- une fonction d'*affectation*, qui précise comment les individus forment les classes autour d'un ensemble de noyaux déterminé ;
- une fonction de *représentation*, qui, à partir d'une partition, propose un ensemble de (nouveaux) noyaux représentatifs ;
- une mesure de la *qualité* de la partition, compte tenu d'un ensemble de noyaux représentatifs.

Ce sont déjà trois points d'entrée offerts pour adapter l'algorithme.

c) *Comparaison avec les méthodes de la même famille*

Les centres mobiles, de Forgy

L'ensemble I à classer est muni d'une distance euclidienne, et, le nombre de classes étant fixé, on mesure la qualité d'une partition par la somme des inerties des classes par rapport à leur centre de gravité. A chaque étape, l'algorithme :

- détermine les centres de gravité des classes de la partition obtenue à l'étape précédente,
- construit une nouvelle partition en agglomérant les éléments autour de ces centres de gravité : la classe j contient les éléments plus proches du $j^{\text{ième}}$ centre de gravité que des autres ; si un élément est à distance minimum de plus d'un centre de gravité, on choisit arbitrairement l'un d'eux,
- continue ou s'arrête selon que la qualité de la partition obtenue est meilleure ou égale à la qualité de la partition obtenue à l'étape précédente.

La partition de départ est choisie par l'analyste : il s'agit d'une partition présumée satisfaisante ou, à défaut, d'une partition quelconque (tirée au sort ou construite par agglomération autour de points tirés au sort).

On peut retenir les différences suivantes, qui sont plutôt à l'avantage des nuées dynamiques :

- définition plus restrictive de la distance, pour pouvoir raisonner en termes d'inertie et de centre de gravité (une distance euclidienne vérifie davantage de propriétés qu'un indice de dissimilarité) ; cette condition permet toutefois d'établir la convergence de l'algorithme. Si l'on a des données qualitatives, une analyse factorielle permet d'obtenir des coordonnées et de définir ainsi une distance ; mais l'analyse factorielle est une procédure relativement lourde.
- centres ponctuels, et ne correspondant pas (du moins pas nécessairement) à des individus.
- critère local pour ajuster les centres des classes : dans les nuées dynamiques, la fonction d'agrégation-écartement peut contrôler que les classes sont bien compactes mais aussi bien séparées.

La méthode de Jancey

Jancey propose le même algorithme que Forgy ; afin de diminuer les risques de « piégeage » dans des optimums locaux, il suggère une variante consistant à prendre pour nouveaux centres d'agrégation, non pas les centres de gravité des dernières classes construites, mais les symétriques par rapport à ces points des centres d'agrégation précédents.

L'intérêt de cette proposition est de rappeler la qualité toute relative des partitions obtenues, et de proposer une piste pour améliorer ce point. Cette piste en tant que telle n'est pas directement transposable aux nuées dynamiques, puisque le calcul du symétrique ne peut s'opérer avec un simple indice de dissimilarité, et que le symétrique est une position, où ne se trouve pas nécessairement un individu.

Les *k-means* de Mac Queen

Dans sa version la plus simple, l'algorithme procède ainsi, étant donné k , le nombre de classes souhaité :

- on prend comme centres d'agrégation les k premiers éléments qui se présentent et on constitue une classe avec chacun d'eux,
- on prend parmi les $n-k$ autres éléments le premier qui se présente et on le réunit au centre le plus proche. On remplace ce dernier par le centre de gravité des deux points réunis. On prend parmi les $n-k-1$ autres éléments le premier qui se présente et on l'affecte au centre le plus proche. On remplace ce dernier par le centre de gravité de la classe formée... et ainsi de suite jusqu'à ce que le dernier élément soit affecté.
On obtient ainsi une partition P .
- On agglomère les points autour des centres de gravité des classes de P , ce qui fournit directement la partition finale P' .

A la différence de la méthode de Forgy, les centres sont donc recalculés après l'affectation de chaque point. Les deux premières phases visent à construire une bonne partition de départ, qui n'est améliorée qu'une fois, dans la dernière phase ; dans la méthode de Forgy, on partait d'une partition sommaire et on l'améliorait autant de fois qu'il était possible.

Cette méthode est la plus rapide de la famille.

d) *Atouts*

La méthode se distingue par ses performances, en terme de temps de calcul et d'espace mémoire requis :

- Comme on n'utilise pas à chaque étape toutes les distances entre toutes les paires d'individus, mais seulement les distances des individus aux noyaux, l'algorithme est approprié au traitement de grandes populations.
- La convergence est rapide : la pratique montre que la partition est généralement obtenue en moins d'une dizaine d'itérations.

La méthode est robuste :

- Les itérations permettent de reprendre et de rectifier les traitements, de fait un peu grossiers en raison de leur rapidité (toutes les distances ne sont pas examinées).
- Si l'on peut toujours obtenir une classification, on a également des indicateurs pour évaluer sa qualité.

La méthode est souple et s'adapte aux données, outre leur grand nombre éventuel :

- L'information apportée par un indice de dissimilarité peut être directement exploitée, sans avoir besoin de recourir à une distance.
- Le cas échéant, on a la possibilité d'exploiter la connaissance que l'on a de la population à classer en fixant les noyaux initiaux.

Les nuées dynamiques s'appuient de plus sur l'idée forte de noyaux :

- Si les individus qui composent le noyau sont bien choisis, ils sont représentatifs, « typiques » de la classe, et en forment un résumé plus riche et plus tangible que peut l'être un centre de gravité.
- Des contraintes peuvent être imposées aux noyaux, dont les éléments par exemple peuvent être choisis parmi des éléments particuliers de l'ensemble initial.
- L'interprétation des résultats peut être facilitée, et leur présentation allégée.
- Les noyaux se moulent en quelque sorte sur la forme de la classe ; l'existence de classes non connexes est un cas particulier qui présente plus de difficulté.

Certes, la méthode ne sollicite pas l'indication d'un seuil arbitraire pour la formation des classes, ou la mise au point d'une heuristique de coupure d'un arbre hiérarchique ; mais elle demande que soit fixé à l'avance un ordre de grandeur du nombre de classes à obtenir.

e) *Points faibles*

La contingence du résultat par rapport à un certain nombre de paramètres initiaux, plus ou moins arbitraires :

- La partition obtenue à l'issue d'un tirage dépend de l'ensemble de noyaux de départ : chaque solution est un optimum local, on ne trouve pas nécessairement un optimum global. La notion de forme forte est une première réponse face à la dispersion possible des résultats, mais la manière d'utiliser les formes fortes vaut d'être approfondie.
- L'algorithme requiert de fixer le nombre de noyaux, qui correspond au nombre de classes obtenues après convergence. Il peut y avoir éventuellement quelques classes vides, c'est pourquoi on recommande généralement de prendre une valeur plutôt légèrement trop grande que trop petite² ; mais il faut tout de même partir d'un bon ordre de grandeur. C'est introduire un *a priori* sur la classification à obtenir, alors que l'on pourrait souhaiter que la partition se façonne directement, et en quelque sorte objectivement, à partir des données. L'issue se dessine encore du côté des formes fortes, dont le nombre peut varier très largement autour du nombre de noyaux³. Il reste qu'un trop grand nombre de noyaux disperse les résultats, et les noyaux se gênent mutuellement. Et un trop petit nombre de noyaux ne laisse pas apparaître des différences et des oppositions significatives.

La méthode est une heuristique, elle n'est pas déterministe et ne fournit pas toujours une solution de la meilleure qualité :

- La convergence n'est démontrée que dans certaines conditions ; en dehors de ces cas, il faut prévoir un traitement assez robuste, qui détecte les cas de non convergence, et qui puisse dans tous les cas proposer une solution acceptable (non aberrante).
- La pluralité des solutions se résout d'un point de vue global, par la recherche de formes stables, – les formes fortes. Cela réduit, sans l'éliminer toutefois, l'incertitude quant à la validité de la classification obtenue. Mais les données réelles s'organisent-elles toujours de manière univoque ? On peut d'ailleurs montrer qu'il y a une instabilité inhérente à la démarche même de classification (Benzécri & al. 1973a, §B.4.3.2). La méthode est peut-être de ce point de vue plus ouverte que des algorithmes qui trouvent par principe une solution et une seule.

4. Mention des méthodes alternatives envisagées les plus intéressantes

a) *L'algorithme de Van Den Driessche*

cf. (Caillez, Pagès 1976, §XV.331)

Présentation

On se donne deux fonctions :

- une fonction *distance*, qui mesure de l'éloignement entre deux classes (ce n'est pas nécessairement une distance au sens mathématique du terme, ni même un indice de dissimilarité) : Van Den Driessche propose de prendre la moyenne arithmétique des dissimilarités entre les éléments des deux classes ;
- une fonction *diamètre*, qui peut être par exemple la moyenne arithmétique des dissimilarités entre tous les éléments de la classe deux à deux.

L'algorithme procède alors comme suit :

² Diday présente ce paramètre comme « le nombre maximum de classes désirées ». Il précise que « quand K est trop grand par rapport au nombre de classes qui existent effectivement, des classes vides apparaissent. »

³ Un calcul simple montre que le nombre de formes fortes se situe en théorie entre $E(N/K^T)$ et $\min(N/2, K^T)$, avec les notations précisées dans les conventions ou définies pour les variables globales de l'algorithme. Par exemple, si l'on classe 100 individus, que l'on se donne 10 noyaux, et que l'on procède à 5 tirages successifs, on peut trouver entre 0 et 50 formes fortes.

Pour amorcer une classe, sélectionner parmi les éléments non encore classés les deux plus proches, puis leur agglomérer celui des éléments non encore classés qui provoque la plus faible augmentation du diamètre, etc. Continuer d'agglomérer tant que les propriétés suivantes sont vérifiées : (i) le diamètre de la classe est inférieur à toute distance (dissimilarité) entre un élément de la classe et un élément non encore classé ; (ii) la distance entre la classe considérée et une classe déjà formée est supérieure aux diamètres des deux classes.

Discussion

Cet algorithme est simple et clair : on « voit » bien ce que traduisent la distance et le diamètre, et on comprend comment ils interviennent.

Il a cependant certains défauts par rapport aux Nuées Dynamiques. En ce qui concerne les performances, il semble un peu plus coûteux car il n'a pas la notion de *noyau*, qui évite de prendre en compte tous les éléments dans les calculs de distance à une classe. Il n'a pas non plus les *formes fortes*, pour tempérer le résultat (déterministe) obtenu. Au cœur de l'algorithme, chaque classe est constituée l'une après l'autre, et l'arrêt d'une classe pour passer à la suivante est net, quelquefois brutal : cela peut intervenir trop tôt ou trop tard. Enfin, l'extrême simplicité de l'algorithme laisse en fait peu de possibilités de réaménagements, en vue de s'orienter vers une classification multiclassées non exhaustive.

b) Deux classifications ascendantes hiérarchiques expérimentalement éprouvées et appréciées

Rappel : raisons du rejet des CAH

Aucune classification ascendante hiérarchique ne fournit une classification multiclassée. Les principaux aspects qui nous ont fait écarter ces algorithmes au profit des Nuées Dynamiques reprennent la plupart des reproches classiques (Caillez, Pagès 1976, §XV.325) :

- leur inadaptation aux grands nombres d'éléments à classer (coût en calculs) ;
- leur déterminisme arbitraire (l'algorithme classe toute population, même s'il n'existe pas de partition « naturelle » ; de plus, si, à une étape, un choix est à faire entre plusieurs fusions, le résultat final dépend presque toujours du choix fait) ;
- la difficulté à trouver un moyen satisfaisant pour déterminer, dans la suite de partitions emboîtées que constitue un arbre de classification, la classification à retenir ;
- la rigidité de l'algorithme : le formalisme est extrêmement précis ; des réaménagements (en vue de permettre la formation de classes chevauchantes par exemple) seraient très délicats à concevoir et à mettre en place.

Utiliser une CAH

Pour autant, les classifications ascendantes hiérarchiques sont une des formes les plus abouties (et les plus pratiquées) des procédures de classification.

Les diverses classifications ascendantes hiérarchiques se déclinent suivant le choix de la fonction de calcul de la distance entre deux classes, elle-même basée sur la fonction de calcul de la distance entre deux éléments. Ces choix influent sur la forme et les propriétés des classes qui seront reconnues : par exemple, le *single-linkage* (distance d'une classe à l'autre = distance minimale entre deux éléments) tend à former des classes « filiformes », globalement peu homogènes. Il importe donc de bien connaître les « biais » propres à chaque méthode, pour choisir celle qui semble la mieux correspondre au contexte d'application (Bommier 1993, p. 31).

Pour notre problème, deux méthodes présenteraient plus de chances de convenir, compte tenu des recommandations des ouvrages de référence et des enseignements tirés de l'expérience.

Méthode de Ward & distance du χ^2

Heuristiquement, la méthode de Ward s'avère souvent supérieure aux autres (Benzécri 1973a, §B.4.2.6) (Saporta 1990, §12.3.4.), nous l'avons nous même observé dans nos travaux (Bommier 1993) et dans ceux de l'équipe : (Hébrail, Marsais 1993), (Assadi 1998), etc.

La méthode de Ward requiert une distance euclidienne ; le χ^2 est une distance souvent choisie (elle exprime l'équivalence distributionnelle). On peut aussi travailler à partir des coordonnées fournies par une analyse factorielle : l'analyse factorielle elle-même demande de gros calculs ; elle permet de réduire la dimension de l'espace et d'estomper les « détails » en ne retenant que les axes les plus significatifs ; Benzécri recommande de normer les coordonnées entre les différents axes retenus (division par la racine carrée de la valeur propre) (Benzécri 1973a, §B.9.2.3).

Average & indice de Jaccard

La distance entre deux classes est donnée par la moyenne (arithmétique) des distances entre un élément d'une classe et un élément de l'autre. C'est un bon compromis entre le *single-linkage* (distance minimale) et le *complete-linkage* (distance maximale).

Cette méthode est mieux adaptée que la précédente sur des tableaux de présence / absence (Benzécri 1973a, §B.4.2.6 & §C.0.3).

c) *Point sur les méthodes disponibles et / ou pratiquées dans l'équipe*

Outils

Les logiciels de classification utilisés par l'équipe sont les procédures de SAS (CLUSTER, peut-être FASTCLUST), pour les classifications ascendantes hiérarchiques, et TEWAT (de la société IBM), qui implémente une analyse relationnelle (le résultat est une partition des individus à classer⁴). Développé en interne au Département TIEM, LEXICLASS obtient également une partition en coupant astucieusement l'arbre obtenu par une classification ascendante hiérarchique (Assadi 1998).

L'équipe a mis au point ses propres outils de classement : CLADOC (analyse discriminante linéaire), et classements de documents par la méthode des k plus proches voisins (dite kNN, *k nearest neighbours*), la proximité entre documents étant fournie par ADOC. Cependant, l'opération de classement (*vs* classification) suppose l'existence préalable de classes, et ne fait qu'indiquer la ou les meilleure(s) classe(s) où affecter un élément non encore classé. Or dans le cas de DECID, il s'agit de former des classes, donc d'un problème de classification (et non de classement).

Conception d'une tactique basée sur les Nuées Dynamiques

La méthode des Nuées Dynamiques a déjà été expérimentée ponctuellement⁵, et s'est montrée satisfaisante. On a cependant observé la nécessité de prendre en compte des cas de non convergence de l'algorithme, avec la formation d'oscillations.

Les Nuées Dynamiques seraient cependant une voie pour obtenir une classification multiclasse non exhaustive, en faisant des *formes fortes* obtenues les noyaux des classes recherchées, puis en procédant à un *classement* multiclasse non exhaustif des éléments restants (qui ne font pas partie d'une forme forte). C'est cette idée⁶ qui a été retenue et développée ici.

⁴ A l'issue d'une classification par TEWAT, les individus forment une partition, mais pas les variables. On a pu ainsi former des classes recouvrantes non exhaustives (en classant les individus et en considérant la structure induite sur les variables, cf. travaux de Richard QUATRAIN, EDF-DER, département SID), mais sans pouvoir s'affranchir de la contrainte de former une partition sur les individus.

Or dans notre cas, ni les unités linguistiques, ni leurs contextes textuels, ne se prêtent à cette structure de partition : une unité linguistique peut être polysémique ou être moins significative que d'autres dans un contexte d'ensemble ; un texte peut comporter plusieurs facettes (sujets, angles d'attaque), peut être original ou général, etc.

⁵ Par Jean-David STA.

⁶ Fruit d'une discussion avec Jean-David STA, EDF-DER, département SID.

B. PRÉSENTATION DU PROGRAMME

1. Conventions

a) Vocabulaire

Les objets que l'on cherche à regrouper sont des *individus*.

On appelle *partie* un ensemble d'individus. On appelle *classe* chacune des parties obtenue après une classification des individus. Les classes forment une partition de l'ensemble des individus de départ, c'est-à-dire que tout individu appartient à une classe et une seule.

Les *noyaux* sont des parties qui servent à construire les classes. Chaque classe a un noyau et un seul, sauf la *classe Z* des individus non affectés. Un *système de noyaux* est un ensemble de noyaux qui décrit une partition, c'est-à-dire qu'il y a une bijection entre l'ensemble des noyaux du système de noyaux et l'ensemble des classes (moins la *classe Z*). Les individus qui constituent un noyau sont appelés *étalons*. Les étalons servent à représenter une classe mais peuvent se trouver classés dans une autre classe : de ce fait, une classe peut se vider. Dans une classe donnée, un individu est soit un étalon, soit un élément *standard* : c'est son *statut* vis-à-vis de cette classe.

Supposons que l'on ait effectué T fois de suite une classification sur l'ensemble des individus. Une *trajectoire* est alors une suite de T classes, telle que la première classe soit issue de la première opération de classification, la deuxième classe de la deuxième classification, et ainsi de suite. Chaque classe fait donc partie d'une des partitions en résultat, et les classes sont données dans l'ordre des classifications successives. Il y a *superposition* de deux trajectoires au niveau d'une partition quand les trajectoires passent par la même classe, que cette classe soit celle des non affectés ou non. On définit l'*écart* entre deux trajectoires comme le nombre des partitions où les trajectoires ne sont pas superposées. Cet écart est donc un entier entre 0 et T , nombre de tirages.

Toute partie regroupant les individus qui ont la même trajectoire est par définition une *forme forte*. Le *degré de variation* des formes fortes est un entier compris entre 0 et 9. Il spécifie la tolérance accordée pour favoriser le regroupement de formes fortes qui auraient été séparées par les aléas propres à l'heuristique des nuées dynamiques. Un degré de variation de valeur i signifie que, si le nombre de classifications effectuées est de 10, alors deux formes fortes dont les trajectoires ont un écart inférieur ou égal à i peuvent être regroupées. Ce degré de variation peut être spécifié *libre* ou *forcé*. Si le degré de variation des formes fortes est libre, alors le programme l'adopte si il convient aux données ; en revanche, si le programme le trouve trop fort, le programme détermine et applique un degré de variation moindre. Un degré de variation forcé permet de forcer le programme à appliquer le degré de variation voulu.

b) Notations

- $\langle xxx \rangle$ est la désignation d'une entité du type xxx . Par exemple, $\langle classe \rangle$ représente la désignation d'un objet qui est une classe.
- Pour x réel, $E(x)$ est la partie entière de x .

2. Entrées / sorties

a) Paramètres de la ligne de commande

Le seul paramètre toujours obligatoire est le **nom du fichier de données** (distances entre individus) et le **nom du fichier des résultats**. Toutes les options sont toujours facultatives, sauf l'option K.

- L'**option K** spécifie un entier strictement positif qui représente le nombre de noyaux. Ce paramètre est obligatoire sauf en présence de l'option M.

- L'**option M** spécifie la distance maximale entre deux individus. Cette distance n'est pas nécessairement atteinte pour les données considérées, en revanche elle n'est jamais dépassée. L'option M ne peut être utilisée que dans le cas où l'on a affaire à une distance bornée. Dans ce cas, l'usage de l'option est recommandé, car le traitement utilise cette information pour ajuster le traitement. Elle donne aussi la possibilité d'alléger le fichier de données.
- L'**option V** spécifie le degré de variation **forcé** pour le regroupement des formes fortes. Elle ne peut être utilisée en même temps que l'option W.
- L'**option W** spécifie le degré de variation **libre** pour le regroupement des formes fortes. Elle ne peut être utilisée en même temps que l'option V.
- L'**option A** indique que l'on souhaite un fichier résultats détaillé. Le fichier de sortie comprend alors des indications sur le déroulement du traitement, permettant d'analyser ce qui a conduit au résultat final.⁷

b) Format et contraintes sur le fichier de données

Le fichier de distances donné en entrée fournit en fait les valeurs d'un indice de dissimilarité. Cet aspect mathématique étant précisé, nous utiliserons par la suite indifféremment dissimilarité ou distance.

Chacune des lignes a le format suivant :
<individu>@<individu>@<dissimilarité>

Par définition, l'indice de dissimilarité entre un individu et lui-même est toujours nul. Les lignes correspondantes peuvent être omises dans le fichier.

Par définition encore, un indice de dissimilarité est une fonction symétrique. La conséquence est la suivante pour le fichier : si on a la ligne A@B@d alors on a soit B@A@d (même valeur de d), soit rien (la distance de B à A est prise comme étant celle de A à B).

Si l'indice de dissimilarité est borné, alors on peut indiquer ce majorant via l'option M (en ligne de commande). Par définition, la valeur fournie pour le majorant doit être supérieure ou égale à toutes les distances dans le fichier. L'utilisation de l'option M permet d'alléger le fichier d'entrée. En effet, dans ce cas, s'il n'y a aucune indication de distance entre deux individus, alors leur distance est considérée comme égale au majorant.

Inversement, si l'on n'utilise pas l'option M, toutes les distances entre deux individus distincts doivent être données. Autrement dit, si les individus A et B apparaissent dans le fichier de distances, alors on doit trouver une ligne commençant par A@B@... ou par B@A@....

c) Format du fichier résultats

Le statut d'un élément dans une classe est noté soit E (pour étalon), soit S (pour standard).

L'homogénéité d'une classe est sa contribution au calcul de la qualité brute (cf. algorithme) : plus la valeur est proche de zéro, meilleure est l'homogénéité de la classe.

Le dernier champ de chaque ligne est un commentaire, qui rappelle à quoi correspond l'information apportée par la ligne.

La présentation des résultats donnée ci-après correspond à une sortie détaillée (option A⁸). On a mis en italiques les lignes qui ne figurent pas dans les résultats standard. En somme, ce que les résultats détaillés ont en plus des résultats standard, c'est la composition et l'évaluation de toutes les classifications intermédiaires avant la classification finale. Soit donc, en plus des lignes commençant par C@R, H@R et Q@R –R comme résultat final–, toutes les autres lignes commençant par C, H et Q –le numéro à la place de R est le numéro du tirage (ordre chronologique). En outre, l'option A affiche à l'écran la version du logiciel, le rappel des noms des fichiers de données et de résultats, et le nombre de noyaux.

⁷ Dans l'implémentation actuelle, l'option A s'appelle option *v*, comme il est d'usage en informatique (*v* pour *verbose*, c'est-à-dire mode détaillé, développé).

⁸ Pour le logiciel implémenté, option *v*.

La ligne commençant par M ne figure que si l'option M a été utilisée. La ligne commençant par V figure sauf quand une variation forcée a été imposée avec l'option V (il n'y a alors pas de degré de variation calculé).

On pourrait aussi mentionner le poids calculé pour l'ensemble initial (quand l'option M est utilisée), et le cas échéant rappeler la valeur du degré de variation indiquée par l'utilisateur avec l'option V ou l'option W.

```
N @ <entier naturel> @ Nombre d'individus
M @ <r el strictement positif> @ Distance par d efaut
T @ <entier strictement positif> @ Nombre de tirages
E @ <r el strictement positif> @ Epsilon (pr ecision du test de convergence)
K @ <entier strictement positif> @ Nombre de noyaux
F @ <entier naturel> @ Nombre de formes fortes
G @ <entier naturel> @ Nombre de classes finalement obtenues
V @ <entier entre 0 et 9> @ Degr e de variation calcul e, utilis e si inf erieur  a W
```

```
H @ 1 @ 1 @ <r el positif> @ Homog en eit e de la classe
C @ 1 @ 1 @ <individu> @ <statut>
C @ 1 @ 1 @ <individu> @ <statut>
C @ 1 @ 1 @ ...
H @ 1 @ 2 @ <r el positif> @ Homog en eit e de la classe
C @ 1 @ 2 @ <individu> @ <statut>
C @ 1 @ 2 @ <individu> @ <statut>
C @ 1 @ 2 @ ...
H @ 1 @ 3 @ <r el positif> @ Homog en eit e de la classe
C @ 1 @ 3 @ <individu> @ <statut>
...
Q @ 1 @ <r el positif> @ Qualit e de la partition obtenue
```

```
H @ 2 @ 1 @ <r el positif> @ Homog en eit e de la classe
C @ 2 @ 1 @ <individu> @ <statut>
C @ 2 @ 1 @ <individu> @ <statut>
C @ 2 @ 1 @ ...
H @ 2 @ 2 @ <r el positif> @ Homog en eit e de la classe
C @ 2 @ 2 @ <individu> @ <statut>
C @ 2 @ 2 @ <individu> @ <statut>
C @ 2 @ 2 @ ...
H @ 2 @ 3 @ <r el positif> @ Homog en eit e de la classe
C @ 2 @ 3 @ <individu> @ <statut>
...
Q @ 2 @ <r el positif> @ Qualit e de la partition obtenue
```

```
...
H @ <valeur de T> @ 1 @ <r el positif> @ Homog en eit e de la classe
C @ <valeur de T> @ 1 @ <individu> @ <statut>
C @ <valeur de T> @ 1 @ <individu> @ <statut>
C @ <valeur de T> @ 1 @ ...
H @ <valeur de T> @ 2 @ <r el positif> @ Homog en eit e de la classe
C @ <valeur de T> @ 2 @ <individu> @ <statut>
C @ <valeur de T> @ 2 @ <individu> @ <statut>
C @ <valeur de T> @ 2 @ ...
H @ <valeur de T> @ 3 @ <r el positif> @ Homog en eit e de la classe
C @ <valeur de T> @ 3 @ <individu> @ <statut>
...
Q @ <valeur de T> @ <r el positif> @ Qualit e de la partition obtenue
```

```
H @ R @ 1 @ <r el positif> @ Homog en eit e de la classe
C @ R @ 1 @ <individu> @ <statut>
C @ R @ 1 @ <individu> @ <statut>
C @ R @ 1 @ ...
H @ R @ 2 @ <r el positif> @ Homog en eit e de la classe
C @ R @ 2 @ <individu> @ <statut>
C @ R @ 2 @ <individu> @ <statut>
C @ R @ 2 @ ...
H @ R @ 3 @ <r el positif> @ Homog en eit e de la classe
C @ R @ 3 @ <individu> @ <statut>
...
```

H @ R @ <valeur de G> @ <r el positif> @ Homog en it  de la classe
C @ R @ <valeur de G> @ <individu> @ <statut>
C @ R @ <valeur de G> @ <individu> @ <statut>
C @ R @ <valeur de G> @ ...
Q @ R @ <r el positif> @ Qualit  de la partition obtenue

3. Disponibilit 

Le programme ici d crit a  t  impl ment  en Ada par Pascal OBRY, EDF-DER d partement SID.

C. ALGORITHME

1. Ressources

a) Constantes et variables globales

Grandeurs fixées par le programmeur :

- T : nombre de tirages (entier strictement positif). On peut fixer T à 10 par exemple.
- $Epsilon$: précision du test de convergence (réel strictement positif). A fixer en fonction de la précision de la machine.
- $Nombre_Maximum_D_Itérations$: sert à arrêter (brutalement) la recherche de noyaux en cas de non convergence⁹.

Grandeurs déterminées par les paramètres en ligne de commande :

- M : distance maximale entre deux individus. Cette distance n'est définie que si elle a été donnée par l'option M .
- Une variable d'état spécifie si le degré de variation est : *indiqué et libre*, ou bien *indiqué et forcé*, ou bien *non indiqué*. Si le degré de variation est indiqué, il est enregistré dans la variable $Degré_De_Variation_Indiqué$.

Grandeurs déterminées au cours du traitement :

- N : nombre d'individus. Il est connu suite à la lecture du fichier de distances.
- K : nombre de noyaux. Il est indiqué par l'option K . Sinon, K est égal au poids de l'ensemble de départ (le poids d'une partie est donné par la fonction P définie ci-après).
- F : nombre de formes fortes. Il est calculé après la série des T tirages.

L'absence de formes fortes ($F = 0$) est le symptôme soit que le nombre de noyaux K est trop grand, soit que les données ne se prêtent pas à être regroupées parce que ses éléments n'ont pas d'affinités particulières vis-à-vis les uns des autres.

b) Fonctions

- $d(<individu>, <individu>)$, distance entre deux individus
Les valeurs de d sont connues par lecture du fichier d'entrée (sachant que d est symétrique), et sinon soit les individus sont identiques, et alors la distance est nulle, soit les individus sont distincts et l'option M est utilisée, et alors la distance vaut M .
- $D(<individu>, <partie>)$, distance d'un individu à une partie :
Nous reprenons la distance proposée par Diday, à savoir :

$$D(i, <partie>) = \sum_{j \in <partie>} d(i, j)$$

- L'écart entre deux ensembles de trajectoires :
L'écart entre deux ensembles de trajectoires est l'écart maximal entre une trajectoire du premier ensemble et une trajectoire du second. L'écart entre deux trajectoires a été défini dans la présentation du vocabulaire. (On s'assure ainsi que toute fusion d'ensemble de trajectoires ne génère pas la réunion de deux trajectoires dont l'écart est supérieur à la tolérance accordée.)
- $Taille_Noyau (<partie>)$, la taille du noyau représentatif d'une partie :
La taille du noyau n'est en fait calculée que pour des classes. On peut prendre la fonction constante égale à 1, c'est-à-dire que le noyau d'une classe est son individu le plus représentatif.
- $Génération_D_Un_Noyau$
Le noyau retourné est le singleton constitué par un individu tiré au hasard parmi les candidats-étalon. L'ensemble des candidats-étalon est l'ensemble de tous les individus qui ne sont pas

⁹ C'est en fait une autre tactique, par « amortissement », qui a été implémentée. Voir description et commentaire dans la partie *Discussion*, paragraphe *Convergence*.

étalons. Dans le cas où l'option M est activée, et qu'en plus la classe Z est non vide, alors l'ensemble des candidats-étalon est la classe Z¹⁰.

- $R(<individu>, <classe>, <système\ de\ noyaux>)$, agrégation-écartement :

La fonction agrégation-écartement sert à qualifier un individu comme représentant d'une classe (représentativité interne, intraclasse : c'est l'agrégation), dans le contexte d'un ensemble de noyaux (caractérisation externe, interclasses : c'est l'écartement). Diday propose deux fonctions :

$$R1(<individu>, <classe>, <ensemble_de_noyaux>) =$$

$$\frac{D(<individu>, noyau_de(<classe>)) \times D(<individu>, <classe>)}{\left(\sum_{<noyau> \in <ensemble_de_noyaux>} D(<individu>, <noyau>) \right)^2}$$

$$R2(<individu>, <classe>, <ensemble_de_noyaux>) = D(<individu>, <classe>)$$

- Déformation ($<système\ de\ noyaux>, <système\ de\ noyaux>$), mesure inverse de la qualité d'une partition :

La déformation mesure la réalisation du critère d'agrégation-écartement, par un système de noyaux localement représentatifs des classes (premier argument) en fonction d'un système de noyaux de référence, constituant un contexte global (second argument). Plus la déformation est petite, meilleure est la qualité de la partition obtenue.

$$\text{Déformation}(\text{Système_Local}, \text{Système_Global}) =$$

$$\sum_{<noyau> \in \text{Système_Local}} \sum_{<étalon> \in <noyau>} R(<étalon>, \text{Classe_De}(<noyau>), \text{Système_Global})$$

Les fonctions ci-après sont définies quand l'option M est utilisée (distance bornée).

- H , le diamètre d'une partie :

Le diamètre de la partie $<partie>$ est la somme des distances de chaque élément de $<partie>$ à chaque élément de $<partie>$.

$$H(<partie>) = \sum_{i \in <partie>} \sum_{j \in <partie>} d(i, j)$$

- P , le poids d'une partie :

Le poids d'une partie est un entier entre 1 et N (le nombre d'individus de l'ensemble initial) : plus précisément, le poids d'une partie ne dépasse jamais le nombre d'éléments contenus dans la partie. Soit n le cardinal de la partie $<partie>$. Si n vaut zéro (partie vide), alors son poids est égal à 1. Sinon :

$$P(<partie>) = E\left(\frac{n^2 M}{n^2 M - H(<partie>)}\right)$$

2. Traitement

a) Série initiale de classifications

Boucle de variation des tirages (parcourue T fois) :

K fois, faire *Génération_D_Un_Noyau* ; constituer à partir de cela le système de noyaux *Système_Courant*. Initialiser le *Nombre_D_Itérations* à 1.

Boucle de convergence à partir d'un tirage des noyaux initiaux :

1. Affectation des individus :

Pour chaque $<individu>$ ¹¹,

¹⁰ Ce dernier cas n'est pas implémenté dans la version actuelle du logiciel.

pour chaque <noyau> du *Système_Courant*, calculer $D(<individu>, <noyau>)$. Si l'option M est active et que <individu> est à distance maximale (M) de chacun des étalons de <noyau>, alors ne pas retenir <noyau> comme affectation possible de <individu>¹².

Rq. : en profiter pour enregistrer la contribution à la *Qualité_Brute* (cf. 2), quand <individu> est un étalon.

Si la valeur minimale de $D(<individu>, <noyau>)$ est atteinte :

- pour un seul noyau : affecter <individu> à ce noyau.
- pour plusieurs noyaux (les « noyaux minimaux ») : affecter <individu> à la classe correspondant à un noyau pris au hasard dans les noyaux minimaux.
- pour aucun noyau : ce n'est possible qu'avec l'option M, et c'est alors le cas de figure d'un individu à distance maximale de tous les étalons. Alors, classer <individu> dans la *classe Z*.

2. Qualité brute :

enregistrement de $Qualité_Brute = Déformation(Système_Courant, Système_Courant)$.

Si le *Nombre_D_Itérations* vaut 1 ou sinon si la *Qualité_Brute* est strictement inférieure à la *Déformation_Minimale*,

enregistrer, comme *Solution_De_Secours*, du *Système_Courant* et de la partition associée, et donner à *Déformation_Minimale* la valeur de la *Qualité_Brute*.

3. Représentation des classes

Pour chaque <classe> non vide,

détermination d'un nouveau noyau constitué de *Taille_Noyau* (<classe>) individus. Pour ce faire, on sélectionne les <individu> non déjà sélectionnés pour un autre nouveau noyau, et minimisant $R(<individu>, <classe>, Système_Courant)$. Les éventuels ex-æquos en surnombre sont départagés par tirage au sort¹³.

Rq. (d'après Diday) : comme on recherche des valeurs minimales de R , si l'on prend $R = R2$ par exemple, le calcul, étant une somme de termes positifs, peut être écourté pour un certain nombre d'individus.

Rq. : en profiter pour enregistrer la contribution à la *Qualité_Ajustée* (cf. 4), pour les individus sélectionnés pour les nouveaux noyaux.

Pour chaque <classe> vide,

lui attribuer un nouveau noyau par *Génération_D_Un_Noyau*.

Rq. : si l'on a choisi une fonction R qui n'est pas nulle lorsque la classe qui lui est passée en paramètre est vide (donc notamment différente de $R1$ et de $R2$), en profiter pour enregistrer la contribution à la *Qualité_Ajustée* (cf. 4), pour les individus sélectionnés pour les nouveaux noyaux.

On dispose maintenant d'un système de nouveaux noyaux, *Système_Nouveau*, à côté du *Système_Courant*.

4. Qualité ajustée

enregistrement de $Qualité_Ajustée = Déformation(Système_Nouveau, Système_Courant)$.

(La qualité ajustée diffère donc de la qualité brute par le fait que les étalons considérés sont ceux du *Système_Nouveau*, alors que le calcul de l'agrégation-écartement se fait toujours en référence au système de noyaux précédent, le *Système_Courant*.)

5. Test de la stabilisation

Si le *Nombre_D_Itérations* atteint le *Nombre_Maximum_D_Itérations*,

on enregistre en mémoire, comme solution pour le tirage, la *Solution_De_Secours*.

Sinon, on calcule :

¹¹ L'algorithme implémenté considère que chaque étalon est d'office affecté à la classe dont il forme le noyau. Il ne reste plus qu'à considérer tous les autres individus non encore affectés.

¹² Cette condition, qui ne joue que lorsque l'option M est utilisée, n'est actuellement pas implémentée.

¹³ La version actuellement implémentée ne procède pas à un tirage aléatoire mais retient le(s) premier(s) des meilleurs éléments trouvés. La principale différence est que c'est alors un processus déterministe.

$$\text{Variation}(\text{Qualité_Brute}, \text{Qualité_Ajustée}) = \left| 1 - \frac{\text{Qualité_Ajustée}}{\text{Qualité_Brute}} \right|$$

Si cette variation est significative, c'est-à-dire supérieure à *Epsilon*,
le *Système_Nouveau* devient le *Système_Courant*,
on incrémente d'une unité le *Nombre_D_Itérations*
et on continue en reprenant la boucle.

Sinon,

on enregistre la partition en mémoire, à savoir le *Système_Courant* et les <classe> associées : c'est la solution produite par le tirage.

Attention, les classes ont gardé leur composition déterminée par la dernière affectation, même si depuis un de leurs éléments a été élu pour devenir un étalon d'une autre classe ; de même, la *classe Z* des non affectés (si *M*) a encore tous ses éléments, même ceux qui ont été sélectionnés pour constituer de nouveaux noyaux.

b) Détermination des formes fortes

On considère la trajectoire de chaque <individu> non encore traité :

si cette trajectoire ne passe jamais par la classe *Z*,

on sélectionne l'ensemble des individus qui ont aussi cette trajectoire (c'est-à-dire les individus qui sont dans la même classe que l'individu considéré, pour tous les tirages). S'il y a effectivement d'autres individus qui suivent la même trajectoire, alors cela construit une forme forte.

Les formes fortes constituent des classes. Leur nombre est enregistré dans la variable *F*. Les individus qui n'appartiennent pas à une forme forte (soit qu'ils aient une trajectoire singulière, soit que leur trajectoire passe par la *classe Z*) constituent la nouvelle *classe Z* des non affectés. Toutes ces classes forment une partition.

c) Degré de variation des formes fortes

Degré_De_Variation désigne le degré de variation des formes fortes rapporté au nombre de tirages.

Si le degré de variation est *indiqué et libre* ou *indiqué et forcé*,

alors on initialise *Degré_De_Variation* à :

$$\text{Degré_De_Variation} = E \left(\frac{\text{Degré_De_Variation_Indiqué}}{10} \times T \right)$$

Si le degré de variation n'est pas *indiqué et forcé*,

on initialise une variable *Degré_De_Variation_Calculé* à zéro ;

on initialise une variable *Test* à zéro ;

Pour un *Degré_A_Tester* croissant à partir de zéro et si nécessaire jusqu'à *T*,

on ajoute à *Test* la valeur

$$\frac{C_T^{\text{Degré_A_Tester}} (K-1)^{\text{Degré_A_Tester}}}{K^T}$$

(le premier facteur du numérateur est une combinaison au sens mathématique du terme, c'est le nombre de parties à *Degré_A_Tester* éléments dans un ensemble à *T* éléments. Pour le calcul, on peut avoir préalablement stocké ces *T+1* valeurs si *T* est fixé par le programme ; on les aura déterminées par un triangle de Pascal. Par exemple, pour 10 : 1, 10, 45, 120, 210, 252, 210, ..., 1).

Si $F \times \text{Test} \geq 1$, alors on arrête de parcourir les *Degré_A_Tester* ;

Sinon,

on donne à *Degré_De_Variation_Calculé* la valeur de *Degré_A_Tester*,

et l'on poursuit la série de tests avec le *Degré_A_Tester* suivant s'il existe.

Les deux cas de figure sont alors les suivants :

1. le degré de variation est *indiqué et libre* :

on remplace la valeur de *Degré_De_Variation* par celle de *Degré_De_Variation_Calculé* si cette dernière est plus petite.

2. Le degré de variation est *non indiqué* :

on donne à *Degré_De_Variation* la valeur de *Degré_De_Variation_Calculé*.

d) Regroupement des formes fortes

Chaque forme forte est une classe, regroupant tous les individus qui ont une certaine trajectoire. On s'intéresse à fusionner certaines formes fortes, de façon à former des classes d'individus ayant des trajectoires dont les écarts peuvent être négligés.

Pour mener à bien ces fusions, on propose un algorithme pour lequel les classes sont dotées d'un statut qui peut prendre trois valeurs : *ouverte*, *en attente*, *fermée*. Au départ, chaque forme forte est une classe *ouverte*, et on cherche peu à peu à agglomérer ces classes jusqu'à ce que toutes les classes soient *fermées*.

Par un léger abus de langage, on parle ici de l'écart entre deux formes fortes pour désigner l'écart entre les trajectoires associées à ces deux formes fortes.

On se donne une variable *Degré_A_Tester* que l'on fixe initialement à 1.

1. [Lancement de l'exploration du voisinage d'une classe]

On part d'une classe *ouverte* quelconque ;

pour chacune de ses formes fortes,

on sélectionne l'ensemble des autres classes *ouvertes* qui ont au moins une forme forte à écart *Degré_A_Tester*.

2. [Bifurcation : poursuite de l'exploration si la classe a un voisinage, passage à la suite sinon]

S'il n'y a aucune classe de la sorte,

la classe considérée est mise *en attente* (on change son statut de *ouverte* à *en attente*) ; on reprend en **4**.

Sinon, on poursuit en **3**.

3. [Recensement exhaustif de tout le voisinage, fusion si le degré de variation est respecté, et passage à la suite]

Pour chacune des formes fortes des classes sélectionnées,

on sélectionne l'ensemble des autres classes *ouvertes* (non encore sélectionnées) qui ont au moins une forme forte à écart *Degré_A_Tester* (au plus) ;

etc., jusqu'à ce qu'il n'y ait plus de nouvelles classes sélectionnées ou qu'il ne reste plus de classe *ouverte*.

On considère alors l'ensemble de toutes les classes sélectionnées (c'est une sorte de single linkage au niveau *Degré_A_Tester*).

Si on trouve une paire de classes dont l'écart est strictement supérieur à *Degré_De_Variation*,

alors toutes les classes (la classe initiale et toutes les classes sélectionnées) prennent le statut *fermée*.

Sinon (pas de paires de classes dont l'écart soit supérieur au *Degré_De_Variation*),

on fusionne ces classes en une seule, qui prend le statut *en attente*.

Puis on reprend en **4**.

4. [Suite : autre classe, ou autre degré, ou fin]

S'il reste des classes *ouvertes*, on reprend en **1**.

Sinon (il n'y a plus de classe *ouverte*) :

si le *Degré_A_Tester* est égal au *Degré_De_Variation*, ou si toutes les classes sont *fermées*, alors **5** ;

sinon (le *Degré_A_Tester* n'a pas encore atteint le *Degré_De_Variation* et il reste des classes *en attente*),

on augmente d'un degré le *Degré_A_Tester*, on change le statut des classes *en attente* en classes *ouvertes*, et on reprend à **1**.

5. On s'arrête.

e) Classement des individus restants

L'ensemble des classes constituées à partir des formes fortes devient le dernier système de noyaux courants *Système_Courant*.

Chaque individu de la classe *Z* des non affectés (ce sont les individus qui n'ont pas une trajectoire de forme forte) est affecté aux classes telles que la distance entre sa trajectoire et la classe des trajectoires des formes fortes du noyau est inférieure ou égale au *Degré_De_Variation*.

Rq. : un individu peut être affecté à plusieurs classes (c'est là la différence majeure avec la procédure d'affectation, dans les classements correspondants à la série de tirages) ; un individu peut aussi n'être affecté à aucune classe.

On obtient donc :

- (i) un ensemble de classes, non nécessairement disjointes, ayant chacune un noyau constitué d'au moins une forme forte ;
- (ii) et une *classe Z* d'individus non affectés.

D. DISCUSSION

1. Explication des apports à l'algorithme original

a) *Equilibre général*

L'algorithme original des nuées dynamiques correspond à la première partie du traitement, la série initiale de classifications. Les seules modifications apportées à ce stade concernent les particularités liées à l'option M, lorsqu'elle est activée. Elles ne portent pas atteinte à la force de l'heuristique, qui réside dans le nombre réduit de calculs (les étalons court-circuitent la prise en compte de toutes les paires d'individus), et dans l'ajustement (rapide) de la solution au fil des itérations.

L'indication d'un majorant de la distance (via l'option M) permet d'abord de calculer une valeur pour K . Le calcul de cette valeur demande le calcul du diamètre, qui demande lui-même la sommation des distances pour toutes les paires. C'est donc un calcul relativement coûteux, mais qui ne grève pas outre mesure la performance générale car il n'intervient qu'une seule fois pour chaque tirage.

L'option M introduit aussi des tests sur les distances. Elle empêche qu'un individu soit affecté à un noyau pour lequel il est à distance maximale de tous les étalons. L'insertion de ces tests, au niveau d'un calcul qui serait de toutes façons effectué, pèse peu, et surtout prévient des affectations dépourvues de sens et potentiellement déséquilibrantes, sous réserves d'une bonne exploitation de la classe Z . *A minima*, l'utilisation de la classe Z comme réserve de candidats pour la génération de nouveaux noyaux contribue à l'efficacité du traitement en limitant la recherche de tels candidats.

Les apports portent donc davantage sur l'exploitation d'une série de résultats, et notamment sur un traitement un peu moins brutal des formes fortes. La proposition de Diday est de prendre les formes fortes comme la dernière série de noyaux, et de retenir la classification obtenue comme solution. Ceci a l'inconvénient de séparer définitivement les individus de formes fortes très proches, distinctes sur le résultat d'un seul tirage. D'autre part, si l'on suit strictement l'algorithme, le résultat final est une partition, or le but ici est d'obtenir une classification multiclasse non exhaustive.

b) *La fonction diamètre D et la fonction poids P*

La fonction de poids évalue le nombre d'éléments différents qui seraient nécessaires pour bien représenter l'ensemble. C'est en quelque sorte une mesure de la diversité interne, de l'hétérogénéité de l'ensemble, ou encore du nombre de pôles que présente l'ensemble.

La fonction poids est définie de manière à prendre les valeurs qui nous intéressent dans trois cas remarquables :

- l'ensemble est composé d'éléments tous identiques (la distance entre deux éléments quelconques est nulle) : alors le poids vaut 1.
- l'ensemble est totalement hétérogène (la distance entre deux éléments quelconques est maximale et égale à M) : alors le poids est égal au nombre total d'éléments.
- un cas intermédiaire est celui de k groupes d'éléments, de même taille, tels que dans chacun des groupes la distance entre deux éléments quelconques est nulle (groupes parfaitement homogènes), et qu'entre deux éléments appartenant à deux groupes différents la distance soit maximale (groupes parfaitement séparés). Dans ce cas, le poids souhaité est k .

Notons n le nombre total d'éléments de l'ensemble considéré. Le troisième cas englobe en fait les deux précédents : le premier correspond à $k = 1$, le second à $k = n$. Le poids correspond en fait à k . Or la géométrie du troisième cas permet d'écrire :

$$\text{diamètre} = (\text{nombre de paires de groupes}) \times (\text{nombre d'éléments dans un groupe})^2 \times M$$

Soit :

$$H = (k(k-1)) \times \left(\frac{n}{k}\right)^2 \times M$$

On en tire k , c'est-à-dire P (puisque $P = k$), en fonction du diamètre :

$$k = P = \frac{n^2 M}{n^2 M - H}$$

La fraction prend ses valeurs entre 1 et n . La fonction de poids P généralisée est donc tout simplement la partie entière de la fraction, avec H la fonction diamètre appliquée à l'ensemble considéré.

De par son interprétation, la fonction P de poids s'applique tout naturellement au choix du nombre de noyaux, qu'elle contribue à rendre moins arbitraire.

c) *La classe Z des non affectés*

Cette classe apparaît quand est définie une distance maximale (M). Dans ce cas, la *classe Z* est le lieu des individus qui n'ont aucun motif de rattachement à aucune classe. Le système des classes existantes est respecté : on n'encombre pas ni ne dénature une classe par rattachement artificiel d'un ou plusieurs individus. La *classe Z* constitue une réserve d'individus, ceux qui ne sont pas (encore) décrits.

La bonne gestion de cette classe est délicate. Mêlés aux autres, les individus mal représentés pouvaient cependant avoir une contribution bénéfique à l'évolution du système des noyaux, en attirant un noyau sinon redondant avec un autre. La délimitation de la *classe Z* instaure une classe sans noyau, et donc sans incidence aucune sur la redéfinition des noyaux des classes non vides. Il faut réintroduire le traitement des noyaux qui, trop proches les uns des autres, se nuisent mutuellement.

d) *Degré de variation des formes fortes*

Le degré de variation des formes fortes est une grandeur interne au calcul, et dont l'impact est relatif à la configuration d'ensemble des données : l'utilisateur manque de repères pour fixer *a priori* une valeur qui donne des résultats satisfaisants. C'est au système que revient de proposer un ordre de grandeur, en fonction des données.

Une forme forte correspond à une trajectoire (ne passant par aucune *classe Z*), c'est-à-dire au choix d'une des K classes pour chacun des T tirages. Dans les conditions habituelles, toutes les trajectoires théoriquement possibles ne sont pas réalisées et représentées par une forme forte (si le cas se présentait, c'est sans doute que K a été pris trop petit, ou/et que les données sont diffuses, et ne se regroupent pas naturellement en pôles). Le nombre de formes fortes est inférieur au nombre théorique de trajectoires différentes.

L'introduction d'un degré de variation permet à une classe de rassembler tout un secteur autour de chaque trajectoire. On considère que le degré de variation est trop grand, si tous les secteurs autour des trajectoires se rejoignent, et que toutes les trajectoires deviennent possibles et équivalentes.

Pour un degré de variation donné, le nombre de trajectoires auxquelles peut s'élargir une forme forte est déterminé. Si les formes fortes sont assez différentes les unes des autres, elles se répartissent l'espace des trajectoires en zones d'influence distinctes. En revanche, des formes fortes particulièrement proche l'une de l'autre entrent dans la zone d'influence l'une de l'autre. Sachant le nombre de trajectoires auxquelles s'élargit théoriquement chaque forme forte, on s'assure que le nombre total de trajectoires reliables à une forme forte ne dépasse pas le nombre de trajectoires possibles, pour ne pas générer une redondance artificielle et gênante.

Le nombre de trajectoires théorique est : K^T

Le nombre de trajectoires couvertes par les formes fortes dépend du degré de variation. Il faut entendre la « couverture » dans un sens large : une même trajectoire peut être comptée plusieurs fois, si elle entre dans la zone d'influence de plusieurs formes fortes. Sans variation, le nombre de trajectoires couvertes est le nombre de formes fortes F . L'introduction d'un premier degré de variation ajoute le choix d'une modification quelconque à chaque trajectoire, ce qui revient au choix

d'un tirage, et dans ce tirage d'une des $K-1$ autres classes. Le second degré de variation permet lui deux modifications, soit donc une variante sur deux tirages.

Par exemple, pour $T=10$ tirages, le nombre de trajectoires couvertes par les formes fortes évolue comme suit :

Degré de variation :	Nombre de trajectoires :
0	F
1	$F \times [1 + 10 \times (K-1)]$
2	$F \times [1 + 10 (K-1) + 45 (K-1)^2]$
t	$F \times [1 + 10 (K-1) + 45 (K-1)^2 + \dots + C(10,t) \times (K-1)^t]$ où $C(10,t)$ désigne le nombre de parties à t éléments d'un ensemble de 10 éléments.
10	$F \times [1 + 10 (K-1) + 45 (K-1)^2 + \dots + C(10,t) \times (K-1)^t + \dots + K^{10}]$ $= F \times [1 + (K-1)]^{10} = F \times K^{10}$ autrement dit, chaque forme forte est élargie à l'ensemble des trajectoires théoriques.

S'agissant d'un ordre de grandeur, pour K pas trop petit (par exemple supérieur à 10), on peut simplifier en considérant que $K-1$ est proche de K . La condition traduisant que le nombre de trajectoires couvertes n'excède pas le nombre de trajectoires théoriques s'exprime alors comme suit :

$$\frac{F \times \sum_{t=0}^{\text{Degré_De_Variation}} C_T^t (K-1)^t}{K^T} \leq 1$$

$$\cong F \times \sum_{t=0}^{\text{Degré_De_Variation}} \frac{C_T^t}{K^{T-t}} \leq 1 \quad (\text{si } K > 10)$$

La somme, qui ne comporte que des termes positifs, est une fonction croissante du *Degré_De_Variation*, tous les autres paramètres étant fixés. Or pour une session donnée, T est fixé (c'est constante du programme), K est fixé (il a été indiqué ou calculé initialement), F est déterminé (nombre de formes fortes obtenues) : l'inéquation peut donc servir à trouver un majorant pour le *Degré_De_Variation*.

e) Regroupement des formes fortes

(Volle 1985) présente la méthode des connexités descendantes pour structurer l'ensemble des formes fortes. Il conclue en pointant l'inconvénient de cette méthode, qui regroupe des formes fortes dès qu'il existe une série de formes fortes intermédiaires pour passer de l'une à l'autre :

le procédé peut fort bien agréger très tôt deux classes qui ont entre elles de fortes différences, mais qui n'ont qu'une faible différence avec une même classe [...]. Ceci est à comparer avec les types d'agrégats obtenus en utilisant la distance « min » [*single-linkage*]. Cet « effet de chaîne » peut-être considéré comme un défaut de la méthode des connexités descendantes. (Volle 1985, §XVII.3)

Nous reprenons cette idée d'agrégation des formes fortes, mais en adoptant la stratégie inverse, de la distance « max » (*complete-linkage*). On obtient ainsi des classes d'une cohésion interne forte, vérifiant le niveau d'exigence fixé. Pour un seuil de tolérance donné, on garantit que la différence entre deux formes fortes d'une même classe n'excède jamais la valeur indiquée. Il n'y a pas l'effet de dérive inhérent à la méthode des connexités descendantes.

Le but est donc d'avoir des classes de formes fortes qui rassemblent les formes fortes que sépare un écart négligeable, et que ce regroupement soit globalement cohérent. Précisément, une classe ne comporte pas deux formes fortes dont l'écart excède le degré de variation prévu (indiqué ou calculé). Les classes sont régulières, au sens où, pour toute forme forte A, si elle est dans la même classe que la forme forte B, et que l'écart entre A et B est inférieur au seuil de cohésion de la classe, alors toutes les autres formes fortes au même écart de A que B font aussi partie de la même classe. Plus généralement, étant donné un élément, les classes considèrent de la même façon des éléments qui lui sont équidistants : (i) soit la distance est inférieure au seuil de cohésion de la classe, et alors ils intègrent tous la même classe que l'élément ; (ii) soit la distance est supérieure au degré de variation prévu, et alors aucun n'intègre cette classe (ils peuvent alors se répartir dans d'autres classes, et pour

certaines rester isolés) ; (iii) soit on se trouve dans le cas intermédiaire : des éléments pourront intégrer la classe s'ils sont à une distance en deçà du seuil de cohésion d'autres éléments de la classe.

L'algorithme proposé teste progressivement, pour chaque classe, un seuil de cohésion de plus en plus élevé. Le seuil de cohésion contrôle les écarts locaux. Le degré de variation indique la limite à ne pas franchir pour les écarts globaux, sur l'ensemble de la classe. Les classes se fusionnent donc tant que le seuil de cohésion de la classe obtenue ne génère pas un degré de variation excessif.

Les statuts des classes s'interprètent dans les mêmes termes. Une valeur du seuil de cohésion étant donnée, une classe est *fermée* si son seuil de cohésion est strictement inférieur à la valeur donnée (donc inutile d'explorer son voisinage pour cette valeur du seuil) ; la classe est *en attente* si son seuil de cohésion est égal et peut-être supérieur à la valeur donnée (le seuil est acquis pour la valeur considérée, on n'explorera que des voisinages pour une valeur de seuil supérieure) ; la classe est *ouverte* si son seuil de cohésion est indéterminé par rapport à la valeur indiquée (l'indétermination doit être levée en explorant le voisinage à la distance du seuil proposé, et la classe devient alors soit *en attente*, soit *fermée*).

f) Classement final

La dernière série de noyaux sont les classes de formes fortes.

Les conditions de rattachement d'un individu à une classe sont moins fortes que celles pour la fusion de formes fortes, si la distance D entre un individu et une classe n'est pas celle d'un *complete linkage*. Une distance de type *complete linkage* consisterait à prendre pour valeur de D la distance de l'individu considéré à l'individu de la classe qui lui est le plus éloigné. Les distances D proposées jusqu'à présent s'apparentent plutôt à une distance moyenne. Le *complete linkage* a un biais favorisant la formation de classes « rondes », c'est-à-dire fortement groupées autour d'un pôle ponctuel ; il rend difficile le développement de classes un peu plus allongées, agençant plusieurs modalités. C'est pourquoi on lui préfère une distance moyenne.

Le *Degré_De_Variation* fixe le plafond au delà duquel l'individu n'est plus rattaché à la classe. Il doit être assez élevé pour que se développe effectivement le multiclassement. Diday définit la notion d'*individu charnière* : « Ce sont les individus qui suivant les tirages oscillent d'une classe à l'autre ». Un *Degré_De_Variation* de 5 ou 6 permet à un individu charnière d'être affecté aux deux ou trois classes avec lesquelles il est en relation. En effet, sur 10 tirages, l'individu peut être 6 fois sur la trajectoire d'une première forme forte, 4 fois sur la trajectoire d'une deuxième, 4 fois aussi sur la trajectoire d'une troisième, et 2 fois sur celle d'une quatrième. Son écart est respectivement de 4, 6 et 8. Plus le *Degré_De_Variation* est élevé, plus il autorise le rattachement à des classes dont les noyaux n'ont aucune affinité entre eux.

La non exhaustivité est ménagée d'une part par l'existence d'une *classe Z* (un individu en relation avec aucune classe n'est pas affecté arbitrairement à une classe), et par le fait que la tolérance permise par le degré de variation se déploie toujours autour d'un rattachement minimal.

2. Points d'évolution

a) Taille des noyaux

En théorie, le nombre d'étalons retenus pour former le noyau est relatif à chaque classe.

- Fixer la taille des noyaux à 1 est la manière de faire la plus simple ; elle est acceptable à défaut d'une stratégie plus évoluée.
- Diday suggère de prendre un nombre fixe, interprété comme une proportion représentative de la population : « en prenant par exemple

$$Taille_Noyau(< partie >) = cste = \frac{3}{4} \times \frac{N}{K}$$

nous faisons simplement l'hypothèse qu'une population peut être caractérisée par les trois quarts de ses individus. Dans la pratique, on constate que la valeur [de la taille des noyaux], à moins d'être très petite, n'influe pas beaucoup sur les résultats surtout s'il existe effectivement des formes fortes. »

- Si l'option M est activée, alors on peut aussi choisir de prendre comme taille du noyau le poids de la classe (donné par la fonction P), puisque le poids est conçu pour évaluer le nombre d'individus représentatifs.
- La taille du noyau pourrait être fonction non seulement de critères internes à la classe, mais aussi de la configuration d'ensemble, par rapport aux autres classes et aux autres noyaux. La fonction ne serait alors pas simplement une fonction de la partie mais aurait aussi comme paramètre un système de noyaux par exemple.

L'intérêt de ne pas limiter le noyau à un seul individu est d'obtenir des noyaux plus représentatifs de la structure interne des classes. Diday commente ainsi un résultat :

On voit l'intérêt des étalons qui jouent le rôle de « squelette » ou de sorte « d'axe factoriel discret » pour chacune des formes reconnues. En prenant trop peu d'étalons ou seulement le centre de gravité on aurait « arrondi » les formes et il n'aurait donc pas été possible de reconnaître les deux formes allongées.

Cependant, en fixant une taille de noyau constante de plusieurs étalons, on court le risque que cette taille soit trop grande pour certaines classes. Par exemple, si une classe se serait naturellement formée de 4 éléments, l'obliger à chercher 6 étalons pour former son noyau n'a pas de sens –sauf si l'on pose que, compte tenu de l'objectif visé par la classification, une classe n'est pas vraiment valide ou viable en deçà de 6 éléments, et que l'on pénalise ainsi délibérément les classes de taille inférieure, même « naturelles ». En somme, adopter une taille de noyau fixe et strictement supérieure à 1 convient si l'on peut faire l'hypothèse que les données se structurent bien (ou doivent se structurer) en classes de taille relativement régulière, ou en tout cas que la taille choisie pour les noyaux est (ou doit être) toujours plus petite que la taille des classes auxquelles s'attendre.

Dans les dernier cas de figure proposés, la taille du noyau varie d'une classe à l'autre. La fonction $D_Cumulative$ pénalise les classes avec un gros noyau, et inversement favorise les classes avec un petit noyau. Cela peut être intéressant, si la taille des noyaux est à la mesure de l'hétérogénéité de la classe (ce que fait la fonction de poids P) : il faut en effet que les classes hétérogènes disparaissent (fondent ou éclatent). Mais aussi, dans un tel contexte, la fonction $Génération_D_Un_Noyau$ doit être redéfinie pour que l'introduction d'un nouveau noyau ne perturbe pas l'équilibre général à cause de sa taille. C'est tout le processus d'évolution au fil des itérations et de convergence qui doit être réexaminé : par exemple, faire en sorte d'avoir un nombre d'étalons fixe, etc.

b) Distance D

La distance proposée par Diday est :

$$D(i, \langle \text{partie} \rangle) = D_Cumulative(i, \langle \text{partie} \rangle) = \sum_{j \in \langle \text{partie} \rangle} d(i, j)$$

Comme cette distance n'est en fait utilisée que pour situer un individu par rapport à différents noyaux, et que les noyaux sont tous de même taille, le comportement de l'algorithme est équivalent à celui que l'on aurait avec la distance moyenne :

$$D_Moyenne(i, \langle \text{partie} \rangle) = \frac{\sum_{j \in \langle \text{partie} \rangle} d(i, j)}{\text{Cardinal}(\langle \text{partie} \rangle)}$$

c) L'agrégation-écartement R

Fonctions proposées par Diday

Diday propose deux fonctions :

$$R1(\langle \text{individu} \rangle, \langle \text{classe} \rangle, \langle \text{ensemble_de_noyaux} \rangle) = \frac{D(\langle \text{individu} \rangle, \text{noyau_de}(\langle \text{classe} \rangle)) \times D(\langle \text{individu} \rangle, \langle \text{classe} \rangle)}{\left(\sum_{\langle \text{noyau} \rangle \in \langle \text{ensemble_de_noyaux} \rangle} D(\langle \text{individu} \rangle, \langle \text{noyau} \rangle) \right)^2}$$

$$R2(< individu >, < classe >, < ensemble_de_noyaux >) = D(< individu >, < classe >)$$

Diday explique $R1$ comme suit : le premier facteur du numérateur « aura pour effet d'agrèger les étalons » ; le deuxième facteur du numérateur « aura pour effet de ramener les étalons vers le centre de leur classe » ; la somme au dénominateur « aura pour effet d'écarter les [noyaux] entre eux ».

Il manque un avertissement important : $R1$ est manifestement inadaptée au cas où la taille des noyaux des classes est fixée à 1 étalon. En effet, les étalons sont alors figés dès le tirage initial : ils ne peuvent évoluer, puisqu'ils minimisent nécessairement $R1$ en lui donnant la valeur nulle. Ceci anéantit toute la stratégie mise en œuvre par les Nuées Dynamiques.

En comparaison, $R2$ peut étonner par sa pauvreté : cette fonction ne considère pas les autres noyaux, elle n'a donc aucun facteur qui exprime l'écartement (souhaitable) des noyaux. La force de $R2$, c'est que c'est une fonction d'agrégation-écartement pour laquelle la convergence est démontrée (en prenant pour distance $D_{D_Cumulative}$, ou de façon équivalente $D_{Moyenne}$ si les tailles des noyaux sont toutes égales et constantes).

Propositions pour une autre fonction R

L'agrégation est une mesure traduisant une densité d'éléments.

En ce qui concerne l'écartement, il est inopportun de décrire l'écartement comme une distance moyenne, vis-à-vis de tous les noyaux : ce qui joue, c'est la distance au noyau le plus proche. Il ne s'agit pas de maximiser la distance globale à tous les noyaux, mais d'éviter des noyaux très (trop) proches. De plus, cette distinction est d'autant plus sensible que le nombre de noyaux est grand, ce qui risque d'être très souvent notre cas. Aussi verrions-nous le facteur écartement exprimé comme un potentiel répulsif, à courte distance : il prend une valeur négligeable s'il n'y a pas de noyau dans son voisinage immédiat, et une valeur vite prépondérante sinon (valeur prépondérante qui atténue l'effet du facteur d'agrégation).

Une fonction illustrant ces propositions est par exemple :

$$R(< individu >, < classe >, < ensemble_de_noyaux >) = D(< individu >, < classe >) \times \left(1 + \left(\frac{4 \times D(< individu >, < classe >)}{\min_{< noyau > \in < ensemble_de_noyaux >} (D(< individu >, < noyau >))} \right)^n \right)$$

Remarques :

Il importe que les valeurs de la distance D soient comparables, normées, donc choisir par exemple $D_{Moyenne}$.

Si l'on veut un calcul plus « économique », on peut remplacer $D(< individu >, < classe >)$ par $D(< individu >, noyau_de(< classe >))$.

L'insistance sur l'aspect écartement peut être modulée par le choix de la puissance n appliquée à l'indicateur de l'écartement (second terme du second facteur). Une attitude moyenne consiste à prendre $n=1$. Avec $n=2$ ou 3, on renforce l'effet répulsif, à courte distance ; avec $n=1/2$, on l'atténue.

d) La fonction de représentation

L'étape de représentation dont il est question est la construction d'un (nouveau) système de noyaux à partir d'une partition.

L'algorithme classique procède classe après classe, et dans chaque classe sélectionne un à un les étalons constitutifs du noyau. Une manière plus juste est de définir non pas la sélection indépendante des étalons, mais le repérage de la partie de taille $Taille_Noyau(< classe >)$ qui est la plus proche (la plus représentative) de la classe considérée, au sens d'une mesure de proximité entre parties. C'est d'ailleurs de cette manière que (Volle 1985) formule la méthode dans toute sa généralité.

e) *Ex-æquo*

Lors de l'affectation d'un individu à une classe, dans la première phase du traitement, il peut arriver que la distance minimale d'un individu donné aux noyaux soit atteinte pour plusieurs noyaux (les « noyaux minimaux »). Comme il faut trancher (car à ce stade il n'y a pas de multiclassement possible), plusieurs tactiques sont envisageables :

- choix au hasard : on tire au sort un noyau parmi les noyaux minimaux, chacun de ces noyaux ayant une probabilité égale d'être tiré. L'avantage de cette tactique est que les individus « intermédiaires » sont répartis de façon équilibrée entre les classes, ils ne viennent pas massivement pénaliser une classe particulière, qui pourrait par ailleurs regrouper d'autres individus plus spécifiques. Ils sont en attente d'une affectation plus significative lors des itérations suivantes, après recalcul des noyaux.
- choix arbitraire déterministe : on choisit par exemple le noyau qui a le plus petit identifiant ; ainsi, si plusieurs individus ont le même choix de noyaux minimaux, ils ne sont pas séparés. S'il y avait ainsi un petit groupe d'individus homogènes équidistants de plusieurs noyaux, leur affectation à la même classe permet qu'ils participent ensemble, à la mesure de leur nombre, à la redéfinition du noyau. On préserve aussi leur possibilité d'appartenance à la même forme forte.
- critère supplémentaire pour départager les noyaux minimaux : par exemple, les autres individus affectés de façon non équivoque à la classe précisent la teneur de la classe, et peuvent rendre une affectation préférable aux autres. On aide ainsi l'affectation à aller dans le sens le plus favorable. Cependant, il peut rester des cas où ce second critère sélectionne encore plusieurs ex-æquos : on n'élimine pas totalement le recours au hasard ou à l'arbitraire. D'autre part, prendre en compte les classes dans leur entier, et plus simplement leur noyau, est contraire à l'orientation générale de l'algorithme. En effet, les noyaux ont l'intérêt d'éviter d'étendre les calculs à toutes les paires d'individus ; et d'autre part, les itérations permettent d'admettre le passage par des états peu satisfaisants, sans compromettre la qualité du résultat : les regroupements se dessinent progressivement.

f) *Génération d'un autre noyau*

Mieux qu'un tirage aléatoire, cette fonction pourrait choisir des étalons dans des zones denses et mal décrites par les noyaux existants. (Wong, Lane 1983) donne un exemple de mesure de densité appliquée à la classification.

La génération d'un nouveau noyau pourrait utiliser la fonction d'agrégation-écartement avec les paramètres suivants : $R(\langle \text{individu} \rangle, \langle \text{partie formée par les } E(N/K) \text{ plus proches voisins de } \langle \text{individu} \rangle \rangle, \langle \text{ensemble des noyaux déjà définis} \rangle)$. Le premier noyau serait choisi sur le critère de densité seul. Des variations seraient à aménager, d'un tirage sur l'autre, pour que le tirage des noyaux initiaux ne soit pas déterministe et toujours identique. Il serait aussi judicieux de contrebalancer la distance aux $E(N/K)$ plus proches voisins par celle au (premier) plus proche voisin, afin de ne pas pénaliser trop fortement les petites classes.

g) *Gestion de la classe Z*

La classe Z sert de « zone neutre » où placer les éléments atypiques, ne se prêtant pas à être regroupés. Elle joue aussi le rôle d'un « réservoir » pour la génération de nouveaux noyaux : elle peut contenir des éléments que rien ne rapproche des noyaux courants, mais qui pourraient en fait se retrouver dans des classes non encore décrites par les noyaux courants. Il faut donc ménager des processus qui permettent l'élimination des noyaux courants peu satisfaisants ou redondants, au bénéfice de l'introduction de nouveaux noyaux plus utiles, mais encore non pris en compte à cause de leur isolement dans la classe Z.

h) Convergence

Théorie et pratique

Diday démontre que les itérations « affectation des individus aux noyaux / représentation des classes par des nouveaux noyaux » convergent vers un optimum local, c'est-à-dire si la fonction d'agrégation-écartement R est *carrée*, à savoir si la fonction *Déformation* définie à partir de R vérifie :

$$\begin{aligned} \text{Déformation}(\text{Système}_Y, \text{Système}_X) &\leq \text{Déformation}(\text{Système}_X, \text{Système}_X) \\ \Rightarrow \text{Déformation}(\text{Système}_Y, \text{Système}_Y) &\leq \text{Déformation}(\text{Système}_Y, \text{Système}_X) \end{aligned}$$

Cette propriété est assez complexe à établir pour la plupart des fonctions R . Pire encore, « toutes les distances ne permettent pas de choisir R carrée » précise Diday ; mais celui-ci se veut néanmoins rassurant : « Dans la pratique on s'aperçoit que la méthode est généralement convergente même si R n'est pas carrée. »

Dans la pratique, on observe aussi des oscillations. Ce peut être un élément dont l'affectation alterne entre deux classes. Mais l'ampleur de l'oscillation peut aussi être supérieure à deux : les itérations bouclent en décrivant une série de n partitions telles que le premier système de noyaux conduit à la définition du deuxième, le deuxième au troisième, etc. jusqu'au n -ième, qui lui conduit au premier. Ces oscillations, de longueur non bornée et non définie a priori, sont plus difficile à repérer.

Heuristiques pour garantir la robustesse du traitement et garder des résultats acceptables

Une tactique envisageable est de procéder au calcul des nouveaux noyaux séquentiellement, en prenant pour paramètre, dans la fonction R d'agrégation-écartement à minimiser, non pas le système de noyaux courant, mais le système de noyaux en cours de renouvellement. L'inconvénient de cette voie est que cela introduit une dépendance à l'ordre de parcours des noyaux (au moment de leur renouvellement).

Une autre stratégie, fruste mais « mieux que rien » : procéder à 100 tirages, garder les 10 meilleurs (au sens de la mesure de la qualité de la partition obtenue).

Comme l'indique Diday, la fonction *Déformation*, calculée à partir de la fonction R d'agrégation-écartement, donne une indication de la qualité de la partition, et donc, en cas de non convergence, on garde préférentiellement comme solution une partition dont le système de noyaux a minimisé la *Déformation*. « Dans tous les cas on aura une bonne solution en prenant la partition qui donne la plus petite valeur à [*Déformation* (*Système_Des_Noyaux*, *Système_Des_Noyaux*)] car cette quantité est proportionnelle au degré d'agrégation des classes et à leur écartement, si R est bien choisie. Remarquons ici que [*la Déformation*] décroît à chaque itération si R est carrée ». C'est la solution qui est introduite dans l'algorithme ici détaillé.

Le logiciel réalisé implémente encore une autre tactique. Si la convergence n'est pas obtenue naturellement (par stabilisation de la Qualité), un *amortissement* des oscillations force la convergence. Pour ce faire, on calcule (le symbole \leftarrow désignant l'opération d'affectation, au sens informatique) :

$$\begin{aligned} \text{si } \text{Nombre_D_Itérations} = 1, \text{Qualité_Moyenne} &\leftarrow \text{Qualité_Brute} ; \\ \text{si } \text{Nombre_D_Itérations} > 1, \end{aligned}$$

$$\text{Qualité_Moyenne} \leftarrow \frac{(\text{Qualité_Moyenne} \times (\text{Nombre_D_Itérations} - 1)) + \text{Qualité_Brute}}{\text{Nombre_D_Itérations}}$$

Autrement dit, la *Qualité_Moyenne* est la moyenne arithmétique de toutes les valeurs de la *Qualité_Brute* pour un tirage des noyaux initiaux, depuis la première valeur calculée jusqu'à la valeur pour l'itération courante. Dès que le *Nombre_D_Itérations* est strictement supérieur à 1, si le test de stabilisation n'est pas vérifié (c'est-à-dire que *Variation* (*Qualité_Brute*, *Qualité_Ajustée*) est supérieure à *Epsilon*), alors on considère la valeur de *Variation* (*Qualité_Brute*, *Qualité_Moyenne*). Si elle est également supérieure à *Epsilon*, on poursuit les itérations ; sinon, on arrête les itérations et on retient comme résultat du tirage la partition courante. Dans la pratique, cette heuristique s'est montrée efficace pour éviter un bouclage infini dans toutes les situations de non convergence rencontrées.

Par delà toutes ces approximations, la force de l'heuristique est dans sa robustesse globale. Le fait de procéder à un certain nombre de tirages relativise les résultats trouvés à chacun, et permet de

ne pas trop faire peser les solutions les plus mauvaises. De plus, on corrige dans une certaine mesure les aléas malheureux en autorisant un *Degré_De_Variation* pour le regroupement des formes fortes.

E. APPENDICE

1. Préparation des données : choix de l'indice de dissimilarité

Décrire les indices existants

Il existe une bonne dizaine d'indices de similarité¹⁴ classiques (Bommier 1993, §AnxSim). Les indices de dissimilarité s'en dérivent par une transformation mathématique simple (*ibid.*).

Pour notre part, nous cherchons à caractériser la (dis)similarité entre deux unités linguistiques (que nous appellerons *mots*, pour simplifier), sachant qu'elles apparaissent chacune dans certains contextes. L'habitude est de noter :

a = nombre de contextes communs aux deux mots,

b = nombre de contextes où seul le premier apparaît, c = nombre de contextes où seul le second apparaît,

d = nombre de contextes où aucun des deux mots n'apparaît.

Interpréter les indices

Considérer de la même manière les deux mots revient à donner un rôle identique à b et c , soit mathématiquement une fonction symétrique en b et c .

Le choix principal restant est de prendre en compte, ou non, le facteur d .

Celui-ci a pour effet de rapporter l'évaluation de la proximité des mots à la globalité du corpus, et non simplement au sous-corpus formé par l'ensemble des contextes concernés par ces mots. Or notre corpus, s'il répond au critère d'homogénéité selon un point de vue donné (par exemple, il peut s'agir d'une collection bien définie de documents, comme l'ensemble des descriptifs d'activité de la DER d'une année), ne présente pas nécessairement une homogénéité thématique d'ensemble. Chaque thématique a son voisinage, sa zone d'influence, mais c'est artificiel de la confronter à toutes les autres qui se trouvent faire aussi partie du corpus.

D'autre part, le facteur d introduit une dépendance à la taille du corpus ; et il est quasiment impossible d'éviter que soient *privilegiés* soit les mots très rares, soit les mots les plus fréquents.

Dans d'autres contextes, Benzécri conclut aussi à la meilleure représentation par des indices qui ne font pas intervenir d (Benzécri & al. 1973a, §C.2.3)¹⁵.

Sélection d'un indice

La première famille d'indices de similarité classiques qui ne font pas intervenir d , est constituée d'indices de la forme suivante (pour α , β et γ rationnels positifs) :

$$\frac{\alpha a}{\beta a + \gamma(b + c)}$$

Prendre β non nul permet d'avoir un indice borné, ce qui nous intéresse : on écarte donc Kulczynski (1927), qui prend $\alpha = \gamma = 1$, $\beta = 0$.

Il reste alors diverses propositions qui prennent $\alpha = \beta$ (= 1 par exemple) pour avoir un indice borné à valeur entre 0 et 1. Reste γ : il est pris égal à :

¹⁴ Nous laissons de côté les indices de distance sur coordonnées réelles, pour ne considérer ici que ce qui est spécifiquement étudié pour les caractéristiques de type présence / absence.

Un mot cependant sur la distance la plus populaire la distance du χ^2 . Elle s'applique à des coordonnées réelles, mais rien n'empêche, techniquement, de l'utiliser pour des variables booléennes en 0 / 1. Un examen de la formule qui calculerait la distance entre deux mots (ou unités linguistiques) en fonction des contextes dans lesquels ils apparaissent, montre que le χ^2 donnerait une importance prépondérante aux contextes courts, ce qui ne nous semble ni justifié, ni pertinent.

¹⁵ Soulignons que ceci est valable dans notre cas, où d traduit bien l'absence d'une caractéristique, et non la valeur alternative d'un caractère *bivalent* (Benzécri & al. 1973a, §B.2.1).

- 1 pour Jaccard (1908)¹⁶,
- 2 pour Sokal & Sneath,
- ½ pour Dice (1945)¹⁷.

Jaccard exprime une *proportion globale* d'accords, Dice une *densité moyenne* d'accords ; Sokal & Sneath ne se laisse expliquer que comme un renforcement (purement numérique) de l'influence du terme $(b + c)$, pour aller dans le sens de Kulczynski (1927), ce qui nous semble moins intéressant, ne serait-ce que sur un plan interprétatif.

Dice fait donc intervenir une notion de moyenne : le dénominateur de la formule est la moyenne arithmétique de la taille des entités dont on mesure la similarité (en l'occurrence, la moyenne des effectifs de nos deux mots). C'est ici que l'on rejoint la seconde (et dernière) famille d'indices de similarité ne faisant pas intervenir d : ces autres indices sont identiques à celui de Dice sauf qu'ils rapportent a à un autre type de moyenne. Il s'agit de Kulczynski (moyenne harmonique) et Ochiaï (moyenne géométrique)¹⁸. L'indice de Kulczynski a un comportement insatisfaisant, en ce que, pour a fixé, il donne plus semblables deux mots l'un très fréquent l'autre très rare, que deux mots de fréquence proche¹⁹.

Pour notre traitement, nous avons dans un premier temps choisi l'indice de dissimilarité calculé à partir de l'indice de similarité de Dice :

$$\frac{b + c}{(a + b) + (a + c)}$$

Ce qui dans notre cas représente le nombre moyen de contextes non partagés (un seul des deux mots apparaît) rapporté au nombre moyen de contextes où apparaissent les mots considérés²⁰.

Cet indice fait intervenir la notion de moyenne, est moins sévère que la dissimilarité calculée à partir de Jaccard (tout en en étant proche), est moins coûteux en calculs que Ochiaï, et est souvent trouvé satisfaisant expérimentalement.

¹⁶ La formule de Jaccard (1908) est la même que celle de Tanimoto, elle peut donc être désignée sous cet autre nom.

¹⁷ L'indice de Dice est encore appelé indice de Czekanowski (1913), ou indice de Sorensen (1948).

¹⁸ Ochiaï est équivalent au cosinus, si l'on pense les critères de présence / absence en termes de coordonnées (0 ou 1) dans un espace vectoriel.

¹⁹ En effet, la moyenne harmonique vaut : $\frac{1}{2} [(1 / (a+b)) + (1 / (a+c))]$. Or cette quantité est d'autant plus grande que $(a+b)$ et $(a+c)$ s'écartent de la valeur centrale donnée par leur moyenne arithmétique.

²⁰ Penser la formule en divisant le numérateur et le dénominateur par 2.

F. REPÈRES BIBLIOGRAPHIQUES

On trouvera un certain nombre de fiches de lectures sur les ouvrages et articles qui suivent dans (Bommier 1993).

1. Article de référence sur la méthode

DIDAY E. (1971) - « Une nouvelle méthode en classification automatique et reconnaissance des formes : la méthode des nuées dynamiques », *Revue de Statistique Appliquée*, XIX (2), pp.19-33.

2. Présentation et commentaires concernant les Nuées Dynamiques

CAILLET F., PAGÈS J.-P. (1976) - *Introduction à l'analyse des données*, S.M.A.S.H., Paris.

SAPORTA G. (1990) - *Probabilités, Analyse des Données et Statistique*, Technip, Paris.

VOLLE Michel (1985) - *Analyse des données*, Economica, coll. Economie et statistiques avancées, 324 pages.

3. Documents complémentaires : classifications en général ; prolongements possibles

ASSADI Housseem (1998) - *Construction d'ontologies à partir de textes techniques Application aux systèmes documentaires*, Thèse de Doctorat, Université de Paris VI, 19 octobre 1998 ; Note interne EDF-DER (co-auteur : Marie-Luce PICARD), HI-23/98/026, accessibilité libre, 292 pages.

BENZÉCRI J.-P. & al. (1973a) - *L'Analyse des Données*, tome I : *La taxinomie*, Dunod ; rééd. 1984, 643 pages.

BOMMIER Bénédicte (1993) - *Recherche d'une typologie des commandes de la DER par analyse statistique des données textuelles*, Rapport de stage de fin d'études, Ecole Centrale Paris, 2 tomes, 73 + 235 pages.

HÉBRIL Georges, MARSAIS Jérôme (1993) - *Typologie de l'activité du centre de recherche de l'EDF par analyse de ses projets de recherche*, Note interne EDF-DER, 94NO00002, octobre 1993, accessibilité libre, 22 pages.

WONG A., LANE T. (1983) - « A k-th Nearest Neighbour Clustering Procedure », *Journal of the Royal Statistical Society, series B*, 45, pp. 362-368.