

Une méthode de classification ascendante hiérarchique par compromis : hclustcompro

Lise Bellanger*, Arthur Coulon** et Philippe Husi**

*Nantes Université, Laboratoire de Mathématiques Jean Leray UMR CNRS 6629,
2 rue de la Houssinière BP 92208, 44322 Nantes Cedex 03, France
lise.bellanger@univ-nantes.fr,

<http://www.math.sciences.univ-nantes.fr/~bellanger/>

**CNRS/Université de Tours, UMR 7324 CITERES, Laboratoire Archéologie et Territoires,
40 rue James Watt, ActiCampus 1, 37200 Tours, France

Résumé. Les méthodes d'apprentissage semi-supervisé permettent d'utiliser des connaissances a priori pour guider l'algorithme de classification dans la découverte de groupes. Dans ce travail, nous proposons un nouvel algorithme de classification de type ascendante hiérarchique (CAH) prenant en compte deux sources d'information associées aux mêmes objets. Cette méthode appelée CAH par compromis (hclustcompro), permet un compromis entre les hiérarchies obtenues à partir de chaque source prise séparément. Une combinaison convexe des dissimilarités associées à chacune des sources est utilisée pour modifier la mesure de dissimilarité dans l'algorithme CAH classique. Le choix du paramètre de mélange est le point clé de la méthode. Nous proposons une fonction objectif à minimiser basée sur la différence absolue des corrélations entre dissimilarités initiales et distances cophénétiques, ainsi qu'une procédure de rééchantillonnage pour assurer la robustesse du choix du paramètre de mélange. Nous illustrons notre méthode avec des données archéologiques provenant du site d'Angkor Thom au Cambodge.

1 Introduction

Les problèmes de classification ou clustering peuvent être abordés à l'aide d'une grande variété de méthodes, toutes nécessitant des techniques dédiées pour la phase de prétraitement des données. Il existe une littérature abondante sur le sujet de la classification, voir par exemple (Aggarwal et Reddy, 2014; Everitt et al., 2001; Kaufman et Rousseeuw, 2005; Nakache et Confais, 2005). Les deux algorithmes les plus connus sont la classification par partition et la classification hiérarchique. Dans cet article, nous nous concentrons sur la classification ascendante hiérarchique (CAH) dont le principe fondamental consiste à construire une structure de données basée sur un arbre binaire appelé le dendrogramme. Quand il existe une information a priori sur les relations entre les individus (e.g. relation de voisinage spatial, relation d'ordre (en génomique), ...),

l'utilisation de méthodes de classification avec contraintes (Ferligoj et Batagelj, 1982; Legendre et al., 1985; Chavent et al., 2018) ou avec prior knowledge (Ma et Dhavala, 2018) permet d'en tenir compte pour construire les groupes. Quand plusieurs sources de données existent, l'utilisation de méthodes de classification par consensus permet d'agréger l'information (Hulot et al., 2020). A l'origine de la méthode présentée dans cet article, se trouve un problème récurrent chez les archéologues, abordé ici dans le cadre d'un projet interdisciplinaire "ModAThorm" (projet ANR, 2018-2023) Modèle explicatif de la fabrique d'Angkor Thom : archéologie d'une ville capitale disparue. Il s'agit de disposer d'un outil statistique d'aide à la périodisation d'ensembles stratigraphiques (niveaux d'occupation, dépotoirs, destruction d'un bâtiment...) pouvant provenir du même site ou de plusieurs sites d'une même ville, parfois spatialement distant donc sans connexion les eux avec les autres. Le résultat prend la forme d'un diagramme des ensembles périodisés jalonnant l'histoire du site en fonction de la proximité temporelle des ensembles. La méthode de classification développée avait donc initialement été appelée perioclust (Bellanger et al., 2021). Mais après discussion avec Gilbert Saporta (PR Emérite CNAM Paris), il s'est avéré plus juste de parler de CAH par compromis et non avec contraintes. De plus, cette méthode a depuis été mise en œuvre sur la bassin de la Loire moyenne (Husi et al., 2022) ainsi que sur des données médicales (Bellanger et al., 2021). Nous avons donc décidé de changer son nom et opté pour un nom plus générique d'où hclustcompro. Cette CAH par compromis est une procédure d'apprentissage semi-supervisé, conçue pour prendre en compte deux sources d'information associées aux mêmes observations et potentiellement sujettes à des erreurs. Une approche basée sur la distance est adoptée pour modifier la mesure de distance dans l'algorithme CAH classique en utilisant une combinaison convexe des deux matrices de dissimilarités initiales. Le choix du paramètre de mélange est donc le point-clé. Nous définissons un critère de sélection de ce paramètre basé sur les distances cophénétiques, ainsi qu'une procédure de rééchantillonnage pour décider du choix du paramètre de mélange dans la méthode de classification proposée. Cet article est organisé comme suit. Dans la section 2, nous décrivons les méthodes existantes tenant compte d'informations a priori. Dans la section 3, nous présentons l'algorithme proposé. Dans la section 4, nous illustrons notre approche sur un jeu de données archéologiques provenant du site d'Angkor Thom au Cambodge.

2 Bref aperçu des méthodes CAH tenant compte d'informations a priori

Les méthodes d'apprentissage semi-supervisé permettent d'utiliser des connaissances a priori pour guider l'algorithme de classification dans la découverte de groupes. Quand il existe une information sur les relations entre les objets (e.g. relation de voisinage spatial, relation d'ordre (en génomique), ...), l'utilisation de méthodes de classification avec contraintes permet d'en tenir compte pour construire les groupes. Dans le cas de plusieurs sources de données, l'utilisation de méthodes de classification par consensus permet d'agréger l'information. En français, une contrainte définit une règle qui impose un certain comportement. De la même manière, la classification avec contraintes est

une classe d’algorithmes d’apprentissage semi-supervisé qui diffère de son homologue sans contraintes en ce sens que les seuls groupes admissibles sont ceux qui respectent plus ou moins strictement la(es) relation(s). Nous n’évoquons ici que celles appelées “Instance Level constraints” (IL) (Davidson et Basu, 2007; Struyf et Džeroski, 2007), spécifiant des règles sur les objets qui peuvent ou non appartenir au même groupe. Les contraintes de type IL ont été incorporées avec succès à l’algorithme CAH (Davidson et Ravi, 2005). Il existe deux grandes approches : (i) celles dans lesquelles l’algorithme de classification est modifié pour intégrer les contraintes, (ii) celles dans lesquelles seule la dissimilarité est modifiée dans l’algorithme de classification. Dans l’approche basée sur l’intégration des contraintes dans l’algorithme, les méthodes CAH basée sur la formule de Lance et Williams (1967) sont facilement modifiables pour intégrer la contrainte. Les algorithmes de classification avec contrainte temporelle (ou spatiale) doivent indiquer sans ambiguïté quels sont les objets voisins. La solution la plus courante pour la classification avec contrainte de contiguïté est d’utiliser des schémas de connexion simples (voir par exemple Legendre et Legendre (2012); Ferligoj et Batagelj (1982)). Cette approche présente quelques inconvénients : (i) elle peut occasionnellement produire des inversions dans le dendrogramme, sauf dans le cas du critère du diamètre (Ferligoj et Batagelj, 1982), (ii) elle ne considère généralement que les dissimilarités entre objets liés, ce qui peut être trop restrictif dans certains domaines comme l’archéologie, comme nous le verrons plus loin. Une approche basée sur la dissimilarité, adaptée aux contraintes de proximité géographique, proposée par Chavent et al. (2018), consiste à modifier la dissimilarité dans l’algorithme CAH. Les contraintes géographiques sont intégrées à travers deux matrices de dissimilarités et un paramètre de mélange. Cette procédure a l’avantage d’être basée sur la dissimilarité et un critère d’hétérogénéité à minimiser à chaque étape pour construire le dendrogramme. Cependant, elle se fonde sur la stratégie d’agrégation de Ward qui ne convient pas à tous les types d’objets et le choix du paramètre de mélange n’est pas toujours évident. Une autre approche (Ma et Dhavala, 2018), basée sur la dissimilarité, consiste à intégrer les connaissances a priori en combinant deux dissimilarités (celle associée aux données originelles et une distance ultramétrique relative aux connaissances a priori). Pour un nombre de groupes fixé, le paramètre de mélange peut être obtenu en maximisant une mesure de stabilité de la partition telles que l’indice de Davies-Bouldin ou l’indice de Dunn. Cette approche présente les inconvénients suivants : (i) les auteurs ne la présentent que pour la CAH avec lien simple et le cas de l’intégration de connaissance ontologique, (ii) le choix du paramètre et du nombre de groupes se fait simultanément à l’aide d’un critère de stabilité de la partition. Enfin, la CAH par consensus (Hulot et al., 2020) conduit à regrouper un ensemble d’arbres ayant les mêmes feuilles pour créer un arbre consensus. Dans l’arbre consensus, un groupe à la hauteur h contient les objets qui sont dans le même groupe pour tous les arbres à la hauteur h . Le principal avantage de cette méthode est de pouvoir travailler avec plus de deux sources d’information. Cependant, par définition d’un consensus, la méthode construit un arbre correspondant à un accord ou consentement du plus grand nombre ; ce qui n’est pas toujours le but recherché. Dans ce travail, nous proposons une approche CAH, appelée CAH par compromis, pour tenir compte des informations disponibles pour deux sources. Notre approche reprend, en l’adaptant, l’idée présente dans certaines méthodes citées précédemment (Chavent

Une méthode de classification ascendante hiérarchique par compromis : `hclustcompro`

et al., 2018; Ma et Dhavala, 2018) de déterminer une combinaison convexe associée à chacune des sources pour construire le dendrogramme.

3 CAH par compromis : `hclustcompro`

Tout d’abord, rappelons qu’en français, le compromis se définit comme une action qui implique des concessions réciproques. En ce sens, la méthode de classification semi-supervisée `hclustcompro` peut être vue comme un compromis entre deux CAH obtenues à l’aide de deux sources d’information.

3.1 Une approche basée sur la dissimilarité

Considérons un ensemble de n objets et notons $\mathbf{D}_1$ (resp. $\mathbf{D}_2$) la matrice de dissimilarités normalisée¹ $n \times n$ associée à la première (resp. deuxième) source d’information. Comme décrit dans Bellanger et al. (Bellanger et al., 2021), le principe de `hclustcompro` est d’appliquer une méthode CAH à la combinaison convexe suivante :

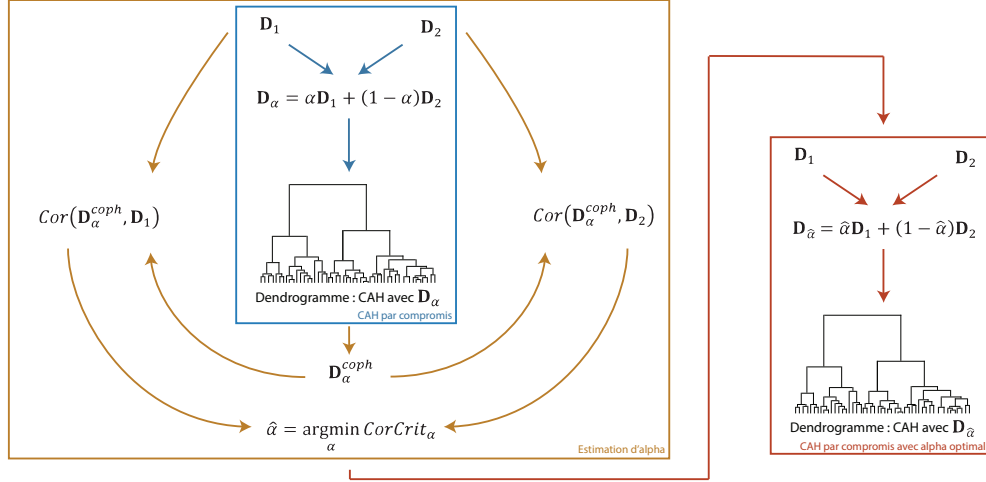
$$\mathbf{D}_\alpha = \alpha \mathbf{D}_1 + (1 - \alpha) \mathbf{D}_2 \quad (1)$$

où $\alpha \in [0; 1]$ est un paramètre fixé qui pondère chaque matrice de dissimilarité (Eq. (1)). Lorsque $\alpha = 0$ (resp. $\alpha = 1$), les dissimilarités obtenues à partir de la matrice de dissimilarités $\mathbf{D}_1$ (resp. $\mathbf{D}_2$) ne sont pas prises en compte dans le processus de classification hiérarchique. Une fois α fixé, le dendrogramme de la CAH peut être construit à l’aide d’une des stratégies d’agrégation satisfaisant la formulation de Lance et Williams. Ainsi, le point clé de cette approche est le choix de α . La détermination de α dépend d’un critère conçu dans l’esprit de la corrélation cophénétique proposée par Sokal et Rohlf (1962). La corrélation cophénétique fait appel à la notion de matrice cophénétique, matrice dont les éléments sont les niveaux de dissimilarité auxquels les objets deviennent membres du même groupe dans le dendrogramme. La corrélation cophénétique correspond à la corrélation linéaire de Pearson entre la matrice de dissimilarité de départ et la matrice cophénétique issue du dendrogramme. Elle permet de mesurer la fidélité avec laquelle un dendrogramme préserve les dissimilarités initiales (voir Sokal et Rohlf (1962); Everitt et al. (2001)). La détermination de α est basée sur l’optimisation de la fonction objectif suivante qui "équilibre" le poids de $\mathbf{D}_1$ et $\mathbf{D}_2$ dans la classification finale :

$$CorCrit_\alpha = |Cor(\mathbf{D}_\alpha^{coph}, \mathbf{D}_1) - Cor(\mathbf{D}_\alpha^{coph}, \mathbf{D}_2)| \quad (2)$$

où $\mathbf{D}_\alpha^{coph}$ est la matrice cophénétique obtenue à partir du dendrogramme issu de la CAH obtenue avec $\mathbf{D}_\alpha$, α fixé dans (Eq. (1)). Le critère $CorCrit_\alpha$ dans (Eq. (2)) représente donc la différence en valeur absolue entre deux corrélations, chacune mesurant la fidélité avec laquelle le dendrogramme obtenu avec $\mathbf{D}_\alpha^{coph}$ préserve les dissimilarités par paire entre les objets initiaux mesurées avec $\mathbf{D}_1$ (resp. $\mathbf{D}_2$). Si deux valeurs sont identiques (par exemple pour $\alpha = 0.5$ et $\alpha = 0.8$) cela signifie que les dendrogrammes produits avec les deux valeurs sont aussi proche de D_1 que de D_2 et par conséquent

1. Les valeurs de dissimilarité sont comprises entre 0 et 1.

FIG. 1 – Schéma illustratif de processus d'estimation de α .

très similaires l'un l'autre et viables. La courbe du critère est plate quand $D_1 = D_2$ (toutes valeurs de α donnant le même résultats). En dehors de ce cas de figure plus les sources initiales sont "différentes" plus la courbe à une forme de "V" avec un intervalle de confiance fin.

La valeur de α est déterminée à l'aide de la formule ci-après :

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \operatorname{CorCrit}_{\alpha} \quad (3)$$

$\hat{\alpha}$ dans (Eq. (3)) peut s'interpréter comme celui conduisant à un dendrogramme représentant la CAH avec $\mathbf{D}_{\alpha}$ défini dans (Eq. (1)) dans lequel la position relative des objets est un compromis entre les dissimilarités $\mathbf{D}_1$ et $\mathbf{D}_2$. La figure 1 ci-dessus illustre le processus d'estimation de α .

Cependant, $\hat{\alpha}$ est une estimation ponctuelle qui ne tient pas compte des erreurs potentielles dans le corpus de données, nous avons proposé une procédure de rééchantillonnage qui permet d'obtenir un intervalle de confiance pour α et d'étudier sa variabilité.

3.2 Stratégie de rééchantillonnage

La stratégie de rééchantillonnage proposée appelée Add One In est conçue dans le même esprit que celle du Leave One Out ; mais elle est basée sur l'ajout d'un "composite" aux objets existants plutôt que sur la suppression d'un objet comme dans le Leave One Out. Une perturbation est effectuée en ajoutant un élément composite. Le composite copie un élément déjà existant, différent pour les deux matrices. Le principe de la méthode est présenté figure 2. Un composite c de l'observation $i \in 1, \dots, n$ est

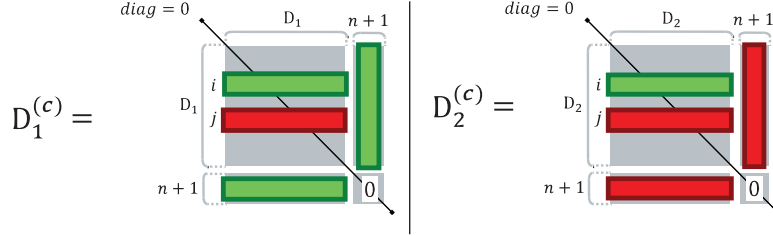


FIG. 2 – Génération d'un composite c pour former $D_1^{(c)}$ et $D_2^{(c)}$. Ajout du composite c copiant l'élément i , dans D_1 , et l'élément i' ($i' \neq i$) dans D_2 .

formé d'une copie de l'observation i de D_1 et d'une copie de l'observation i' ($i' \neq i$) de D_2 soit $n-1$ possibilités. Un total de $n(n-1)$ composites peuvent alors être créés. Les matrices de dissimilarités $D_1^{(c)}$ et $D_2^{(c)}$ de dimension $(n+1) \times (n+1)$ représentent les dissimilarités pour un ensemble d'objets composé des objets originaux et du composite c fixé.

α est estimé à partir de l'éq. (2) en remplaçant D_1 et D_2 par $D_1^{(c)}$ et $D_2^{(c)}$ respectivement. Le paramètre $\hat{\alpha}$ peut être calculé pour chaque composite possible c afin d'estimer l'écart-type et d'obtenir un intervalle de confiance basé sur la méthode des percentiles pour α (voir algorithme 1). Une CAH peut alors être réalisée en utilisant $D_{\hat{\alpha}}$ ou $D_{\tilde{\alpha}}$ où $\tilde{\alpha}$ est au voisinage de $\hat{\alpha}$ dans le respect de l'intervalle de confiance $IC_{95\%}(\alpha)$.

Algorithm 1 Add one in procédure – Ecart-type estimé et $IC_{95\%}(\alpha)$ par la méthode des percentiles pour α .

for $c \in 1, \dots, n(n-1)$ **do**
 — Générer un composite et créer $D_1^{(c)}$ et $D_2^{(c)}$
 — Définir $D_{\alpha}^{(c)}$ et $CorCrit_{\alpha}^{(c)}$ où :
 — $D_{\alpha}^{(c)} = (1-\alpha)D_1^{(c)} + \alpha D_2^{(c)}$ et
 — $CorCrit_{\alpha}^{(c)} = |Cor(D_{\alpha}^{coph(c)}, D_1^{(c)}) - Cor(D_{\alpha}^{coph(c)}, D_2^{(c)})|$
 — Evaluer $\hat{\alpha}^{(c)} = \operatorname{argmin}_{\alpha \in [0;1]} CorCrit_{\alpha}^{(c)}$; réplicat de $\hat{\alpha}$ pour chaque composite c
end for

Obtenir :

— $\hat{\alpha}^* = \frac{1}{n(n-1)} \sum_{c=1}^{n(n-1)} \hat{\alpha}^{(c)}$, estimation ponctuelle de α à partir des $\hat{\alpha}^{(c)}$;
 — $\hat{se}^* = \sqrt{\frac{\sum_{c=1}^{n(n-1)} (\hat{\alpha}^{(c)} - \hat{\alpha}^*)^2}{n(n-1)-1}}$, écart-type estimé de $se(\hat{\alpha}^*)$;
 — $IC_{95\%}(\alpha)$ intervalle de confiance (méthode des percentiles) à 95%.

Dans le cas où le nombre d'objets n est grand, la possibilité de ne pas calculer

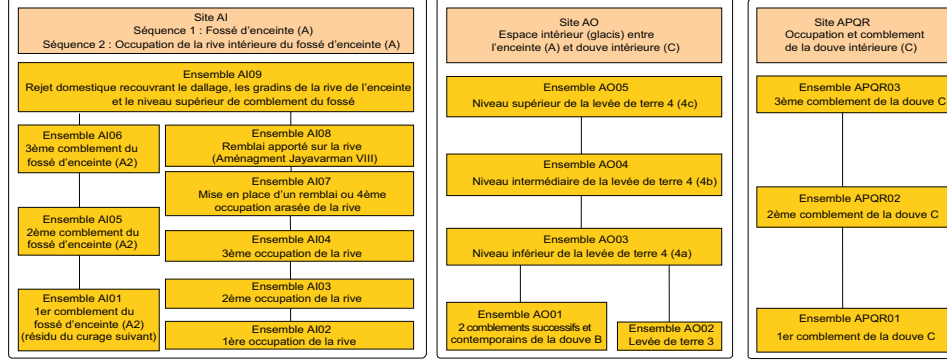


FIG. 3 – *Diagramme stratigraphique de sériation de trois sites archéologiques en relation avec le système d'enceinte de la ville d'Angkor Thom.*

l'intervalle de confiance sur l'ensemble des $n(n-1)$ composites a été développé afin de réduire le temps de calcul. Il est donc possible de choisir un nombre x (fixé plus petit que $n-1$) de composites calculés pour l'observation i . Cette option permet de (i) réduire le nombre de possibilités de $n(n-1)$ composites à $n.x$ composites, (ii) diminuer le temps de calcul à des temps raisonnables et tout en conservant un nombre de composites suffisant. Une CAH peut alors être réalisée en utilisant $\mathbf{D}_{\hat{\alpha}}$ ou $\mathbf{D}_{\tilde{\alpha}}$ où $\tilde{\alpha}$ est proche de $\hat{\alpha}$ et dans l'intervalle de confiance. L'algorithme de CAH par compromis est implémenté dans la fonction `hclustcompro` du package R `SPARTAAS` (Coulon et al., 2023). Ce package dispose d'un ensemble de fonctions permettant de sélectionner un α optimal, couper le dendrogramme pour construire une partition ou subdiviser un groupe. Une version `Shiny` est également disponible pour les utilisateurs occasionnels du logiciel R².

4 Résultats de la CAH par compromis sur des données archéologiques

Dans cette section, nous présentons les résultats obtenus avec la CAH par compromis (`hclustcompro`) sur des données issues de différents sites fouillés à Angkor Thom (Cambodge), capitale de l'empire khmer entre le IXe et le XVe s. (Gaucher, 2004). Nous comparons également les résultats à ceux obtenus avec une CAH classique pour interpréter l'apport de notre méthode dans le cas des données étudiées.

4.1 Les données archéologiques

L'un des objectifs majeurs ici est de préciser la périodisation de la ville, notamment à partir (i) du diagramme de sériation ou stratigraphique (Fig. 3) autrement appelé

2. <http://www.r-project.org>.

Une méthode de classification ascendante hiérarchique par compromis : `hclustcompro`

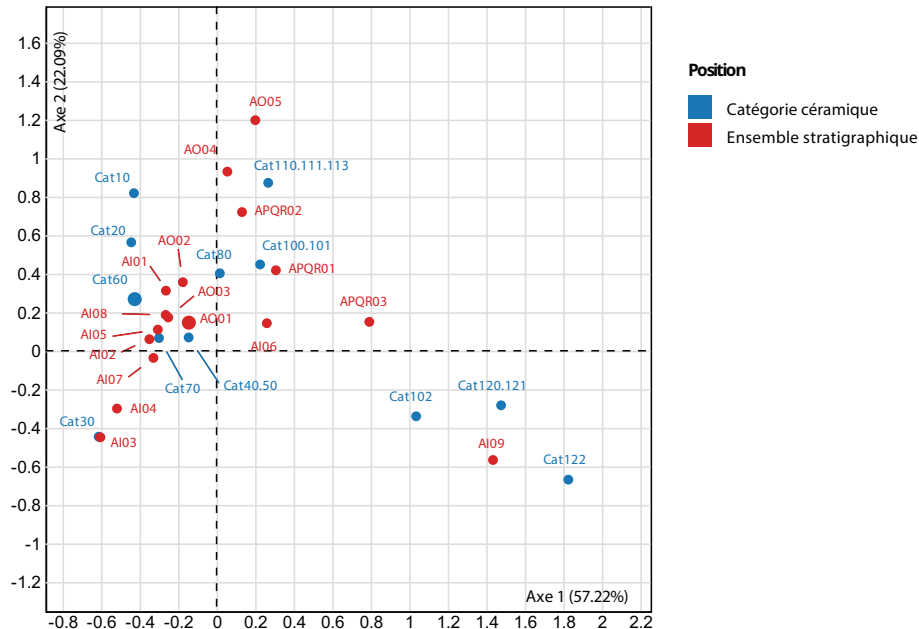


FIG. 4 – *Plan 1-2 de l'AFC sur N.*

«matrice de Harris» (Harris, 1989) illustrant les relations physiques « sur/sous » donc chronologiques «avant/après» entre les 17 ensembles provenant de 3 sites archéologiques séparés (ii) des assemblages céramiques qui leurs sont associés (quantité de tessons par catégorie de céramique et par ensemble stratigraphique). La céramique (vaisselle en terre cuite) a l'avantage d'être indestructible, omniprésente dans les fouilles, avec des changements typologiques rapides dans le temps, ce qui en fait une des meilleures sources de datation en archéologie.

À partir du diagramme de sériation (Fig. 3), et donc des relations de connection des 17 ensembles, il est donc possible de construire $\mathbf{S}_2$, la matrice symétrique d'adjacence définie comme une matrice binaire de connectivité, puis $\mathbf{D}_2 = \mathbf{1}_{17 \times 17} - \mathbf{S}_2$ associée aux 17 ensembles stratigraphiques. L'information sur la céramique est contenue dans une table de contingence $\mathbf{N}$ de taille 17×12 où les lignes correspondent aux ensembles et les colonnes aux catégories céramiques. Comme très souvent sur ce type de données (Bellanger et Husi, 2012), l'analyse factorielle des correspondances (AFC) (Greenacre, 2016) sur $\mathbf{N}$ permet d'observer dans le plan factoriel 1-2 une forme parabolique dite « en fer à cheval » (effet Guttman) des projections des profils-lignes et colonnes (Fig. 4). Cette forme est révélatrice ici d'une évolution chronologique : l'ordre dans lequel se répartissent les catégories et les ensembles présente une séquence évolutive, mais d'autres facteurs peuvent entrer en ligne de compte. La CAH couplée à l'AFC est très souvent utilisée pour définir des groupes d'ensembles. Cependant ces groupes ne tiennent pas compte de l'information sur la stratigraphie.

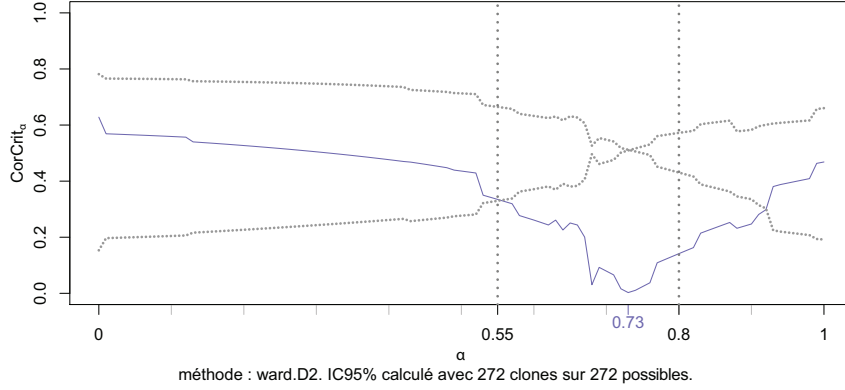


FIG. 5 – $CorCrit_\alpha$ en fonction de α . $\hat{\alpha} = 0.73 \in [0.55 : 0.8]$; $\tilde{\alpha} = 0.7$.

D'où l'idée d'enrichir la construction et l'interprétation globale de la chronologie des 3 sites en combinant l'information sur la céramique à celle découlant du diagramme stratigraphique (Fig. 3) à l'aide d'une méthode de classification adaptée telle que `hclustcompro`. Les distances euclidiennes entre les ensembles sont calculées à partir de toutes les composantes des profils-lignes de l'AFC sur $\mathbf{N}$. Les CAH sur $\mathbf{D}_1$ représentant la céramique et $\mathbf{D}_2$ représentant la stratigraphie conduisent aux valeurs les plus élevées du coefficient d'agglomération (Kaufman et Rousseeuw, 2005) pour le critère de Ward. Elle peut être considérée comme la meilleure stratégie d'agrégation à adopter pour ces données. Les dendrogrammes obtenus séparément à l'aide d'une CAH avec critère de Ward peuvent être comparés à l'aide du coefficient d'entrelacement qui prend des valeurs comprises entre 0 et 1 ; une valeur faible traduisant de très grandes similitudes entre les 2 dendrogrammes. Dans notre cas, l'entrelacement vaut 0.39 : les dendrogrammes sont relativement similaires, mais pas identiques. Cela confirme que les informations fournies par la céramique et la stratigraphie doivent être considérées simultanément pour résoudre le problème de classification.

4.2 Obtention d'une partition avec `hclustcompro`

Pour appliquer `hclustcompro`, nous définissons $\mathbf{D}_\alpha$ à partir de (Eq. (1)) et déterminons un α optimal en utilisant (Eq. (3)) avec un intervalle de confiance issu de la stratégie de rééchantillonnage (voir Sect. 3.2). Nous obtenons $IC_{95\%}(\alpha) = [0.55; 0.80]$ et choisissons $\hat{\alpha} = 0.7$ (Fig. 5). Cette valeur indique que pour les données d'Angkor Thom, le poids de chaque source d'information est réparti comme suit : 70% pour la céramique et 30% pour la stratigraphie. Ce déséquilibre peut résulter d'une stratigraphie dont les limites ne sont pas toujours clairement définies, conséquence de perturbations liées à l'importance des moussons, donc de l'eau au cours du temps.

Une CAH avec critère de Ward est effectuée à partir de la matrice de dissimilarité $\mathbf{D}_{0.7}$ comme défini dans (Eq. (1)). Le nombre de groupes à retenir a été choisi en

Une méthode de classification ascendante hiérarchique par compromis : hclustcompro

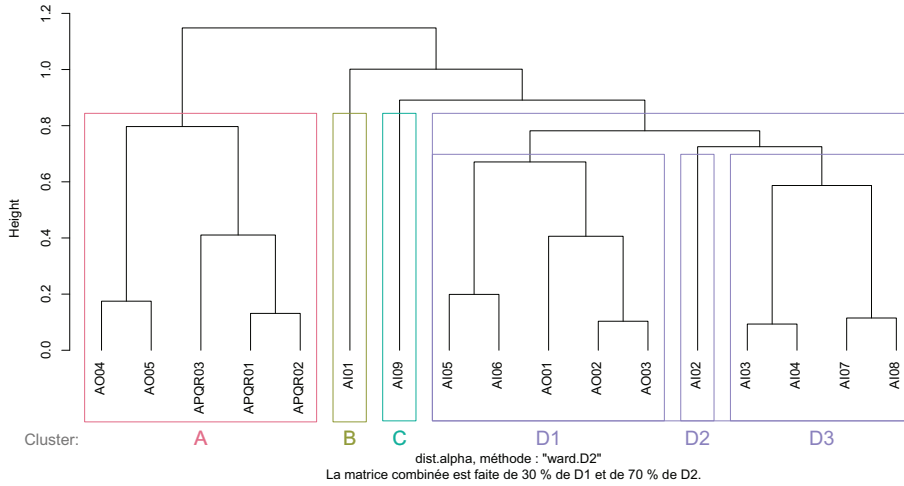


FIG. 6 – Dendrogramme *hclustcompro* ($\alpha = 0.7$) critère de Ward : 4 groupes avec un groupe subdivisé en 3.

fonction de l'examen de l'échelle des indices d'agrégation associés au dendrogramme (Fig. 6) ; mais aussi en fonction de la connaissance du site par l'archéologue. En effet, le choix de 4 groupes avec le groupe D divisé en 3 sous-groupes (Fig. 6) semble le mieux adapté aux rythmes chronologiques de la ville.

4.3 Comparaison entre CAH par compromis et CAH

La figure 7 présente le dendrogramme issu de la CAH classique avec critère de Ward sur les données céramiques. L'absence de prise en compte de la stratigraphie rend ces résultats difficilement interprétables archéologiquement. En effet, des ensembles en relation physique (Fig. 3) situés entre deux autres peuvent se retrouver de manière incohérente dans un autre groupe, comme ici APQR02 isolé de APQR01 et APQR03, cas de figure qui n'existe pas avec *hclustcompro*.

L'indice de Rand ajusté (Rand, 1971) montre une relative similitude (0.69) entre les deux partitions en quatre groupes. On observe malgré tout, des différences intéressantes (voir tableau 1).

Archéologiquement, trois ensembles sont pour certains bien plus anciens (AI01 et AI02) pour d'autres bien plus récents (AI09) que les autres caractérisant les principaux états de construction et d'occupation de l'enceinte. Dans la partition CAH classique, seul l'ensemble AI01 ressort clairement comme isolé des autres. Celle issue de la CAH par compromis traduisant mieux la réalité chronologique de l'histoire du site : elle permet d'identifier AI01, dans une moindre mesure AI02 mais surtout AI09 comme isolé des autres ensembles sachant que ce dernier est bien plus récent que ceux qui le précèdent. La CAH par compromis permet d'intégrer la source d'information stratigra-

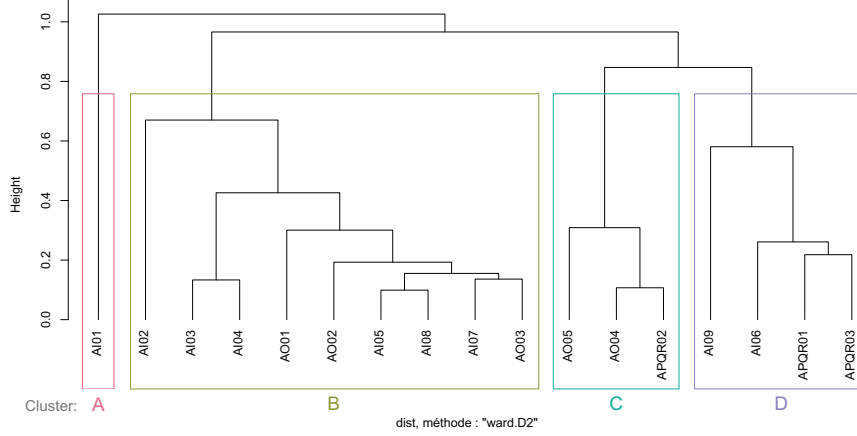


FIG. 7 – Dendrogramme CAH classique critère de Ward : 4 groupes.

	Partitions	hclustcompro			
		A	B	C	D
hclust	A	0	1	0	0
	B	0	0	0	9
	C	3	0	0	0
	D	2	0	1	1

TAB. 1 – Matrice de confusion entre partitions CAH par compromis et CAH classique.

phique pour construire une partition, là où d'autres méthodes de type CAH traiteraient cette information comme une contrainte trop forte.

5 Conclusions

Dans ce travail, nous avons présenté une nouvelle méthode de CAH basée sur un compromis entre deux sources d'information disponibles. Cette approche fondée sur une modification de la dissimilarité dans l'algorithme de CAH classique est simple à mettre en œuvre. La matrice de dissimilarités modifiée dans la CAH est une combinaison linéaire de deux matrices de dissimilarités, donc par construction tous les critères d'agrégation existants peuvent être utilisés. Les problèmes du choix et de l'interprétation du paramètre de mélange, points clés pour ce type de méthodes de classification, sont résolus. Le paramètre de mélange définit l'importance donnée à chaque source dans la procédure de classification. Bien que hclustcompro ait été conçue à l'origine pour répondre à un problème archéologique, cette méthode présente un intérêt dans

Une méthode de classification ascendante hiérarchique par compromis : `hclustcompro`

de nombreux autres domaines d'application comme par exemple la santé (Bellanger et al., 2021). La CAH par compromis trouve une résonnance toute particulière avec les méthodes factorielles d'analyse conjointe de plusieurs tableaux de données telle que STATIS qui recherche un tableau compromis le plus représentatif selon certains critères. Partant de ce constat, la perspective méthodologique naturelle est d'étendre notre méthode au cas de plus de deux sources d'information croisant les mêmes objets.

Remerciements : Cette recherche a été soutenue par le projet ANR ModAThOm coordonné par Philippe Husi et Jacques Gaucher (EFEO). Les auteurs tiennent à remercier Jacques Gaucher pour ses commentaires et son expertise des données et Gilbert Saporta pour nous avoir suggéré la dénomination “CAH par compromis”.

Références

- Coulon, A., L. Bellanger, et P. Husi (2023). Spartaas: Statistical pattern recognition and dating using archaeological artefacts assemblages. R package version 1.2.1, <https://spartaas.gitpages.huma-num.fr/r-package/>.
- Husi, P., V. Aubourg, L. Bellanger, A. Bocquet, J. Bouillon, M. Bonnard, E. Coffineau, A. Coulon, A. Finet, L. Fricot, M. Gary, C. Gerbaud, E. Jaffrot, S. Jesset, D. Josset, C. Marie Lacroix, C. Lejault, C. Leparmentier, A. Longelin, F. Marteaux, S. Millet, I. Moréra-Vingotte, A. Noël, C. Rigault, P. Testard, et B. Véquaud (2022). *La céramique médiévale et moderne du bassin de la Loire moyenne, chronotypologie et transformation des aires culturelles dans la longue durée (6e—17e s.)*. 79e Supplément à la Revue Archéologique du Centre de la France, FERACF, Tours, <https://ceramedvaldeloire.huma-num.fr/editions/suppl79racf2022>.
- Bellanger, L., L. Chevreuil, P. Drouin, D. Laplaud, et A. Stamm (2021). Peut-on détecter des troubles de la marche avant qu'ils ne soient perceptibles? *Revue Tangente Hors-série Bib73 de la "Bibliothèque Tangente" sur Maths et emploi en entreprise*.
- Bellanger, L., A. Coulon, et P. Husi (2021). Perioclust : A simple hierarchical agglomerative clustering approach including constraints. In T. Chadjipadelis, B. Lausen, A. Markos, T. R. Lee, A. Montanari, et R. Nugent (Eds.), *Data Analysis and Rationality in a Complex World*, Cham, pp. 1–8. Springer International Publishing.
- Hulot, A., J. Chiquet, F. Jaffrézic, et G. Rigail (2020). Fast tree aggregation for consensus hierarchical clustering. *BMC Bioinformatics*, <https://doi.org/10.1186/s12859-020-3453-6>.
- Ma, X. et S. Dhavala (2018). Hierarchical clustering with prior knowledge.
- Chavent, M., V. Kuentz-Simonet, A. Labenne, et J. Saracco (2018). Clustgeo: an R package for hierarchical clustering with spatial constraints. *Computational Statistics* 33(4), 1799–1822.

- Greenacre, M. (2016). *Correspondence Analysis in Practice*. Boca Raton: Chapman and Hall/CRC.
- Aggarwal, C. et C. Reddy (2014). *Data Clustering: Algorithms and Applications*. Boca Raton: Chapman and Hall/CRC.
- Legendre, P. et L. Legendre (2012). *Numerical ecology. 3rd ed.* Amsterdam: Elsevier Science BV.
- Davidson, I. et S. Basu (2007). A survey of clustering with instance level. *ACM Transactions on Knowledge Discovery from Data*, 1–41.
- Struyf, J. et S. Džeroski (2007). Clustering trees with instance level constraints. In J. N. Kok, J. Koronacki, R. L. d. Mantaras, S. Matwin, D. Mladenič, et A. Skowron (Eds.), *Machine Learning: ECML 2007*, Berlin, Heidelberg, pp. 359–370. Springer Berlin Heidelberg.
- Kaufman, L. et P. Rousseeuw (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley-Interscience.
- Nakache, J.-P. et J. Confais (2005). *Approche pragmatique de la Classification*. Paris : Editions Technip.
- Davidson, I. et S. S. Ravi (2005). Agglomerative hierarchical clustering with constraints : Theoretical and empirical results. In A. M. Jorge, L. Torgo, P. Brazdil, R. Camacho, et J. Gama (Eds.), *Knowledge Discovery in Databases : PKDD 2005*, Berlin, Heidelberg, pp. 59–70. Springer Berlin Heidelberg.
- Gaucher, J. (2004). Angkor thom, une utopie réalisée ? structuration de l'espace et modèle indien d'urbanisme dans le cambodge ancien. *Arts Asiatiques* 59, 58–86.
- Everitt, B., S. Landau, et L. Morven (2001). *Cluster Analysis. 4th ed.* Oxford: Oxford University Press Inc.
- Harris, E. C. (1989). *Principles of Archaeological Stratigraphy 2nd ed.* London and San Diego: Academic Press.
- Legendre, P., S. Dallot, et L. Legendre (1985). Succession of species within a community: Chronological clustering, with applications to marine and freshwater zooplankton. *The American Naturalist* 125(2), 257–288, doi:10.1086/284340.
- Ferligoj, A. et V. Batagelj (1982). Clustering with relational constraint. *Psychometrika* 47(4), 413–426.
- Rand, W. M. (1971). "objective criteria for the evaluation of clustering methods". *Journal of the American Statistical Association* 66(336), 846–850.
- Lance, G. N. et W. T. Williams (1967). A general theory of classificatory sorting strategies ii. clustering systems. *The computer journal* 10(3), 271–277.
- Sokal, R. R. et F. Rohlf (1962). The comparison of dendrograms by objective methods. *Taxon* XI (2), 33–40.

Summary

Semi-supervised learning methods allow to use a priori knowledge to guide the classification algorithm in group discovery. In this work, we propose a new hierarchical agglomerative clustering algorithm (HAC) that takes into account two sources of information associated with the same objects. This method, called compromise HAC (hclustcompro), allows a compromise between the hierarchies obtained from each source taken separately. A convex combination of the dissimilarities associated with each of the sources is used to modify the dissimilarity measure in the classical HAC algorithm. The choice of the mixing parameter is the key point of the method. We propose an objective function to be minimized based on the absolute difference of correlations between initial dissimilarities and cophenetics distances, as well as a resampling procedure to ensure the robustness of the choice of the mixing parameter. We illustrate our method with archaeological data from the Angkor site in Cambodia.